Portland State University
PDXScholar

Dissertations and Theses | Dissertations and Theses

1985

Use of synthetic speech in tests of speech discrimination

Jane S. Gordon
Portland State University

Let us know how access to this document benefits you.

Follow this and additional works at: http://pdxscholar.library.pdx.edu/open_access_etds

 Part of the Communication Technology and New Media Commons, and the Speech and Hearing Science Commons

Recommended Citation
Gordon, Jane S., "Use of synthetic speech in tests of speech discrimination" (1985). *Dissertations and Theses.* Paper 3443.

This Thesis is brought to you for free and open access. It has been accepted for inclusion in Dissertations and Theses by an authorized administrator of PDXScholar. For more information, please contact pdxscholar@pdx.edu.

AN ABSTRACT OF THE THESIS OF Jane S. Gordon for the Master of Science in Speech Communication presented May 10, 1985.

TITLE: Use of Synthetic Speech in Tests of Speech Discrimination.

APPROVED BY MEMBERS OF THE THESIS COMMITTEE:

[redacted]

James F. Maurer, Chairman

[redacted]

John C. McDermott

[redacted]

Richard J. Sonnen

The purpose of this study was to develop two tape-recorded synthetic speech discrimination test tapes and assess their intelligibility in order to determine whether or not synthetic speech was intelligible and if it would prove useful in speech discrimination testing. Four scramblings of the second NU-6 monosyllable word list were generated by the ECHO II speech synthesizer using two methods of generating synthetic speech called TEXTALKER and SPEAKEASY. These stimuli were presented in one ear to forty normal-hearing adult

subjects, 36 females and 4 males, at 60 dB HL under headphones. Each subject listened to two different scramblings of the 50 monosyllable word list, one scrambling generated by TEXTALKER and the other scrambling generated by SPEAKEASY. The order in which the TEXTALKER and SPEAKEASY mode of presentation occurred as well as which ear to test per subject was randomly determined.

The mean performance scores for TEXTALKER and SPEAKEASY demonstrated that the intelligibility of synthetic speech produced by the ECHO II speech synthesizer was significantly reduced in comparison to the intelligibility of normal human speech and to synthetic speech produced by expensive formant resonance synthesizers. However, the mean performance scores reported in this study compared favorably to those of other researchers who used the more affordable and commercially available speech synthesizers.

The mean performance score for SPEAKEASY was slightly higher than the mean performance score for TEXTALKER but the difference was not statistically significant. There was relatively little learning which occurred when subjects listened to synthetic speech generated by TEXTALKER and SPEAKEASY. Also, there was no effect noticed for the different scramblings of the second NU-6 word list for TEXTALKER and SPEAKEASY. The test-retest reliability appeared to remain relatively stable for TEXTALKER and SPEAKEASY. Several test words such as "young", "room", "match", "rain" and "learn" were highly intelligible while others such as "keg", "shack", "gaze", "goal", "said", "pad" and "shawl" were almost totally unintelligible for both TEXTALKER and SPEAKEASY. The medial phonemes or vowels were most intelligible while

initial and final consonants were approximately equal in intelligibility. The voiceless fricatives /f,θ,ʃ, h/ were very poorly produced. These results suggest that synthetic speech as generated by the ECHO II speech synthesizer using TEXTALKER and SPEAKEASY methods of generation is not of sufficient intelligibility at this time to be used as a substitute for normal human speech in speech discrimination testing.

USE OF SYNTHETIC SPEECH IN TESTS OF SPEECH DISCRIMINATION

by

Jane S. Gordon

A thesis submitted in partial fulfillment of the

requirements for the degree of

MASTER OF SCIENCE
in
SPEECH
with emphasis in Speech Pathology
and Audiology

Portland State University

1985

TO THE OFFICE OF GRADUATE STUDIES AND RESEARCH:

The members of the Committee approve the thesis of

Jane S. Gordon presented on May 10, 1985.

[redacted]

James F. Maurer, Chairman

[redacted]

John C. McDermott

[redacted]

Richard J. Sonnen

APPROVED:

[redacted]

Theodore G. Grove, Head, Department of Speech Communication

[redacted]

Jim F. Heath, Dean of Graduate Studies and Research

ACKNOWLEDGEMENTS

There are several people I would like to thank who without their help and support I would not have been able to complete this research project.

I sincerely appreciate all of the help and encouragement that Dr. James F. Maurer, my thesis committee chairman and graduate advisor, gave me during the completion of this study. His patience and sense of humor through it all was very comforting.

A special thanks goes to Dr. John C. McDermott who gave expert advice on technical details of writing, research design, and statistical analysis. He was very generous with his time and always available to answer questions.

I want to especially thank my dear friends, Sue Ann Manderfeld, Catherine Moore, and Cheryl Comfort, who cared for my two young children while I worked on this research project. Without knowing my children were happy and loved I would not have been able to leave them.

I also wish to thank all those students who particpated as subjects in this study. Their interest and willingness to help was rewarding and greatly appreciated. I would also like to thank Peggy Pryor who helped immensely with her skillful and speedy typing skills.

I dedicate this thesis to my husband, whose constant and reassuring love and support enabled me to see this study through to the

end; to Sarah and Elizabeth who waited patiently for their mother to fulfill a goal; to my parents who taught me "to always finish what I start"; and finally, to my Lord and Savior, Jesus Christ, who gave me spiritual strength in times of need.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF TABLES (continued)

LIST OF TABLES (continued)

LIST OF FIGURES

CHAPTER I

INTRODUCTION

Speech audiometry has been widely recognized as an important addition to pure tone audiometry in the audiological test battery. However, unlike the relatively simple pure tone stimuli, the variables inherent in speech stimuli are more difficult to control and quantify. These variables include the various vocal qualities and articulation characteristics of the speaker, the test administration hearing level, the types of speech materials used in the test such as syllables, words, sentences, or continuous discourse, the speed of message delivery, and the characteristics of the test equipment.

One of the greatest sources of variability is the individual speaker's presentation. A great deal of controversy exists concerning whether speech materials should be presented via monitored live-voice or from a recording. Those critical of the live-voice method argue that the results obtained by different speakers cannot be compared unless it can be demonstrated that the speakers are equivalent. In the past, recorded speech tests have been used in an attempt to standardize the monitored live-voice method. However, because a speaker's unique vocal characteristics are permanently built into a recording, there may be as much difference between two recordings as there is between two live-voice talkers. An example of the variability that can occur

between two recordings is exemplified by the dissimilarity that exists between the more difficult Rush Hughes recording of the Harvard Psycho-Acoustic Laboratories (PAL) PB words and the relatively easy recording of the CID W-22 PB words by Ira Hirsh. Recorded versions of the same test material may also produce different results as evidenced by the findings concerning the various recordings of the NU-6 test.

Since speech audiometry is becoming so important and so widely used, there is a real need for standardization of materials and methods of production and presentation so that results from different clinics and laboratories can be compared. What is needed, is a standardized way to make a consistent recording of speech stimuli that is reproducible from clinic to clinic and that does not contain individual vocal characteristics inherent in various speakers.

Synthetic speech, or the production of speech by a computer with an electronic voice synthesizer attached, offers a new approach to traditional speech audiometry. At present it is the only available means which makes possible the reproduction of fully identical speech test materials. Traditional recordings utilizing natural or human speech are limited by the individual articulation characteristics of the speaker who makes the recording. Standardized synthetic speech would not have this problem and could create a more uniform speech stimulus that would be fully reproducible and comparable between clinics. Test results would be attributable to the test itself and not to variables introduced by the speaker. Synthetic speech may be the only method to create a lasting basis of comparison between speech tests. Even with conventional recordings, master tapes eventually

wear out and have to be re-recorded, thereby sacrificing some validity. However, with synthetic speech, fully identical test tapes can be reproduced over and over again.

Due to the problems inherent in traditional speech testing, using monitored live-voice or recorded methods of presentation, this study will develop two synthetic speech test tapes and assess the intelligibility of synthetic speech for possible use in audiological assessment. This study will try to determine whether or not synthetic speech is intelligible and, if it will prove useful in speech discrimination testing.

CHAPTER II

REVIEW OF THE LITERATURE

SPEECH AUDIOMETRY

Pure tone audiometry is limited in its diagnostic value because it does not directly assess the listener's hearing for speech stimuli. Speech audiometry provides a method by which such an assessment can be made and therefore, has become an increasingly useful addition to pure tone audiometry (Carhart, 1953a; Berger, 1971). By measuring a person's ability to hear in situations that reflect the salient features of everyday listening, speech audiometry increases the validity of a complete audiological assessment and provides important diagnostic and prognostic information for the patient concerning receptive communication skill (Hirsh, et al., 1952; Silverman and Hirsh, 1955).

Speech audiometry usually employs two types of tests (Carhart, 1953a; Ventry, et al., 1971, and Tillman and Olsen, 1973). The first test is composed of spondaic words or disyllabic equal stress words which are very homogenous in audibility and are used to measure auditory sensitivity or threshold sensitivity for speech (Jerger, et al, 1968; Tillman and Olsen, 1973; Goetzinger, 1978). This is usually called the speech reception threshold (SRT). It is defined as the lowest intensity level (dB HL) at which 50 percent of the test items

are correctly identified by the listener (Tillman and Olsen, 1973). In addition to threshold tests, speech signals which are highly heterogeneous in audibility are used to assess suprathreshold intelligibility, or the ability to understand speech at levels above threshold (Carhart, 1965; Ventry, et al., 1971; Tillman and Olsen, 1973). These suprathreshold tests are commonly known as tests of speech intelligibility or as speech discrimination tests with performance scores calculated in percentages (Carhart, 1953a; Ventry, et al., 1971).

Speech discrimination tests are useful in estimating an individual's ability to communicate in social situations (Silverman and Hirsh, 1955; Ventry, et al., 1971), as well as assisting in the determination of site of lesion within the auditory system (Jerger, et al., 1968; Ventry et al., 1971; Goetzinger, 1978), and predicting the outcome of medical treatment and otological surgery. Finally, speech discrimination tests are useful in evaluating the need for and proper selection of hearing aids, and in measuring the progress of aural rehabilitative procedures, such as lipreading and auditory training (Carhart, 1953b; Jerger, et al., 1968; Goetzinger, 1978). The many and diverse applications of speech discrimination testing exemplifies its high degree of usefulness. Speech discrimination testing is useful in every phase of modern audiology, including medical, educational, and research applications (Carhart, 1953b; Jerger, et al., 1968).

Problems Inherent in Tests of Speech Discrimination

Speech intelligibility testing is not without certain problems. Some of the problems encountered with traditional speech discrimination testing include those which are introduced by the personnel involved in the testing situation itself, that is by the speakers and listeners. Another class of variables involved in speech testing are the types of materials used in the test. These include nonsense syllables, words, sentences and continuous discourse. A final variable is the communication equipment used for testing, which includes the rooms, microphones, amplifiers, radios, and earphones, etc. (Egan, 1948; Miller, et al., 1951).

Personnel

Speaker. The speaker brings a host of variables to the speech testing situation due to the various qualities inherent in his or her voice and articulation. Some of the variables involved in speech production include vocal force or intensity, pitch level, rapidity of speech, and overall vocal quality (Brandy, 1966; Goetzinger, 1978). Individual speaker's voices may also vary due to regional dialects, articulation characteristics, and steadiness or variability of speech power (Egan, 1948; Goetzinger, 1978). Consistency of the reader during monitored live-voice testing is a common clinical problem.

A commonly accepted notion is that men's voices are easier for hard of hearing individuals to understand than women's, because women's voices are fainter and higher in frequency (Palmer, 1955; Sanders, 1971). The average habitual pitch level for male speakers is around 120 Hz whereas for females it is around 200 Hz (Sanders, 1971).

However, Palmer (1955) found no difference in intelligibility between male and female voices for hard of hearing individuals. He suggested that if differences were found in the articulation score between different speakers, that the examiner should investigate other articulatory characteristics of the different speakers besides pitch level.

A controversial issue among audiologists concerns whether materials should be presented by monitored live-voice (MLV), that is, spoken by the tester at the time of the test, or presented from a pre-recording. Those critical of the live-voice method argue that the test results obtained by two different speakers cannot be compared unless it can be determined that they are equivalent (Carhart, 1965). Brandy (1966) states that two or more speakers produce variations in listener performance due to talker-by-list and talker-by-distortion interaction. He further states that a single talker's presentation may vary from one day to another, thereby causing a difference in the listener's performance. In other words, the same speaker's voice may vary enough on different days to effect a change in the discrimination score of the listener. However, with the exception of the Veteran's Administration Audiology clinics, live-voice testing is the most popular form of administration, because of its speed of administration.

Administering pre-recorded speech discrimination tests is also not without problems. Because each speaker's unique vocal characteristics are built into the recording, there may be as much difference between two recordings as there is between two live-voice speakers (Carhart, 1965). This fact became evident in the widely

published disparity between the relatively easy recording of the CID W-22 PB word lists by Ira Hirsh and the more difficult recording by Rush Hughes of the Harvard PB-50 word materials (Carhart, 1965; Elkins, 1970; Stevens, 1978). Although these two recordings represent different test word materials, a great part of the difference in their intelligibility was due to the individual characteristics of the two speakers' voices. Goetzinger (1978) states that the articulation of Ira Hirsh on the W-22 recordings is superior to Rush Hughes' presentation on the Harvard recordings. The dissimilarity that can occur between recorded versions of the same test material has been documented regarding various recordings of the NU-6 test. Rintelmann, et al. (1974) found his own recording more difficult than the original recording by Tillman and Carhart. In reference to the variability among different recordings, Goetzinger (1978) states the need for standardized, recorded versions of speech discrimination materials.

Despite the limitations in the use of recorded speech tests, there are many who support the use of recorded speech material over the live-voice method (Silverman and Hirsh, 1955; Brandy, 1966; Hood and Poole, 1980). Brandy (1966) states that recorded presentations are more reliable than live-voice presentations. Hood and Poole (1980) suggest that recorded lists are considered mandatory for standardization, accuracy and repeatability. Tillman and Olsen (1973) state that no standardized test is possible unless recorded lists are used due to the fact that talker differences in monitored live-voice defy standardization. Regardless of whether speech materials are presented by live-voice or from a recording, an appreciable amount of

difficulty in the speech discrimination task is uniquely imparted by the speaker's voice (Hood and Poole, 1980).

Listener. The listener brings his or her own set of unique variables to the testing situation. Listeners differ in their ability to hear speech (Egan, 1948). This variability arises from such attributes as listener intelligence or educational background, attention, and motivation (Hirsh, et al., 1952; Hood and Poole, 1980). The variability in listeners is a factor which examiners have little control and very little knowledge (Hood and Poole, 1980).

The effects of training or learning regarding the task involved are also variables in speech discrimination testing that can influence test scores (Egan, 1948; Hirsh, et al, 1952; Schwartz and Goldman, 1974). Egan (1948) found that practice has an important influence on the speech discrimination score of PB-50 word materials and that articulation scores with inexperienced listeners show improvement with practice under difficult listening conditions. He advised that listeners be trained until little or no improvement is shown in their test scores. Goetzinger (1978) advises using a practice period consisting of approximately five words from another word list when administering recorded PB word tests.

Egan (1948) found that fatigue seems to have little effect on the outcome of test scores, provided that sufficient rest periods are given during the test situation. Ross and Huntington (1962) also support a rest period in their testing design.

Scoring Method. A final variable is brought to the testing situation by the speaker or examiner and the listener. This has to do

with the method of scoring the listener's responses. Responses can be scored in one of two ways, either by having the listener repeat what he hears or write down what he hears. In speech reception threshold testing, where the test words are few in number and easily identified, the examiner is likely to have little difficulty determining whether or not the listener's responses are correct (Berger, 1971). However, in speech discrimination testing, the speech stimuli are not only greater in number but also more difficult to identify. In this situation, the examiner is thus placed in the position of being a listener and the question is raised as to whose speech discrimination is really being tested, the examiner's or the listener's (Jerger, et al., 1968; Berger, 1971). There is good evidence that this factor can lead to unreliable speech discrimination testing, particularly if the examiner has a hearing loss. Nelson and Chaiklin (1970) noted that the speaker is inclined to overestimate the correct responses of the listener in speech discrimination tests that use a talkback scoring procedure. These researchers concluded that the range of this correct scoring bias is sufficient to question the validity concerning the clinical use of talkback responses. An added advantage of the written response method is that it allows for an analysis of specific articulation errors. The "write-down" procedure is a safer method, although it is tedious and time consuming, and limited by the listener's spelling ability and handwriting legibility (Jerger, et al., 1968; Berger, 1971).

Types of Test Material

Speech discrimination tests are influenced by the type of speech stimuli used in testing and their familiarity to the listener

(Goetzinger, 1978). Speech materials consist of nonsense syllables, words, sentences, and continuous discourse, ranging in a hierarchy of difficulty from greatest to least (Miller, et al., 1951; Goetzinger, 1978). Whether or not nonsense syllables, words or sentences are used as the speech stimuli in testing depends upon a number of considerations.

Nonsense Syllables. Nonsense syllables are single syllables composed of meaningless combinations of speech sounds. The use of nonsense syllables gives a more accurate indication of the number of phonemes actually heard by the listener than do tests that are composed of words and sentences. Another advantage in using nonsense syllables as speech stimuli is that syllable lists can be generated with little difficulty (Egan, 1948). On the other hand, nonsense syllables can pose several problems. Their usage requires that the speaker as well as the listener be trained. The speaker must articulate the speech sounds accurately while the listener must be able to record the speech sounds heard using phonetic symbols (Egan, 1948; Miller, et al., 1951). Nonsense syllables are also extremely confusing due to the fact that they are very abstract. Thus they create a rather difficult speech discrimination task for the listener (Carhart, 1965). Using nonsense syllables poses a difficult listening task from another standpoint. The listener is faced with a limitless range of possible alternatives from which to choose. The perception of previous phonemes in the syllable gives no clue as to the ones which follow, and the listener feels that any choice is possible (Miller, et al., 1951).

Monosyllabic Words. Using words as stimuli in a speech

discrimination test overcomes several of the disadvantages of nonsense syllables. Monosyllabic words, which are meaningful words given out of context as isolated units, are more intelligible than nonsense syllables. The intelligibility of a word improves markedly with the increase in the number of sounds within the word (Egan, 1948; Hirsh, et al., 1954). The use of words instead of nonsense syllables as the speech stimuli is not as confusing and thus represents a less difficult listening task (Carhart, 1965; Goetzinger, 1978).

The traditional test for speech discrimination used in the United States consists of word lists composed of 50 phonetically balanced monosyllabic words (Carhart, 1965; Tillman and Olsen, 1973). Egan and his colleagues at the Harvard Psycho-Acoustic Laboratory (PAL) constructed the original word lists during World War II (Carhart, 1965; Hood and Poole, 1980). Egan (1948) outlined several criteria on which the construction of the word lists was based: the words chosen were monosyllabic in structure and in common usage in the English language; the word lists were phonetically balanced, that is, the speech sounds within the lists occur with the same relative frequency as they do in a representative sample of American English speech; the word lists were equal in range and average level of difficulty (Egan, 1948; Campbell, 1965; Carhart, 1965; Hood and Poole, 1980).

The criteria that common, monosyllabic words be chosen in constructing the word lists has been generally accepted without serious question (Campbell, 1965). In constructing a discrimination test for speech, relatively nonredundant items are necessary. Monosyllabic words are chosen in lieu of conversational sentences or multisyllabic

words, because monosyllabic words are sufficiently unpredictable enough that their perception relies solely on the accurate discrimination of individual speech sounds and is not due to a multiplicity of other cues (Carhart, 1965). Familiar words are chosen to minimize the effects of differences in the educational background or level of intelligence of various subjects (Hirsh, et al., 1952; Levin, 1952).

The criteria that the word lists be phonetically balanced has been a subject of considerable investigation among various researchers. As of yet, experimental evidence that phonetic balance is necessary as far as clinical validity is concerned appears to be nonexistent (Campbell, 1965).

The specifications that the word lists be equal in range and average level of difficulty are necessary when constructing any set of reliable and equivalent tests (Campbell, 1965). Word lists must be equal in this regard in order to allow comparison of scores between word lists.

The original word lists at Harvard were developed from a pool of 1200 monosyllabic words according to the above criteria as specified by Egan (1948). The PAL PB-50 word lists consisted of twenty lists each containing 50 words. Rush Hughes, a professional radio announcer from Boston, recorded eight of the word lists (Goetzinger, 1978; Hood and Poole, 1980). The PAL PB-50 word lists were found to be inadequate for several reasons. The lists contained many unfamiliar words and their phonetic balance was questioned. The recordings were poorly standardized and not all of the lists were found to be equivalent (Carhart, 1965; Goetzinger, 1978; Hood and Poole, 1980).

Hirsh, et al. (1952) undertook a modification of the PAL PB-50 word lists and constructed the well-known Central Institute for the Deaf (CID) W-22 test. They selected 120 words from a pool of 1000 PB-50 words and added 80 new words. Subsequently, the 200 words were arranged into four phonetically balanced lists, each containing 50 words and then recorded by Ira Hirsh (Goetzinger, 1978).

The CID W-22 and the Harvard PB-50 word lists were the most widely used word lists for speech discrimination testing during the 1950's and 1960's (Elkins, 1970). However, many differences exist between the two tests and much controversy regarding their usage remains. The Rush Hughes recording of the PAL PB-50 word lists was found to be a much more difficult test than the CID W-22 test (Carhart, 1965; Tillman and Olsen, 1973; Goetzinger, 1978). This can be seen by examining the articulation-vs-gain function produced by the two tests. Figure 1 shows the articulation fucntion for the CID W-22 versus the PAL PB-50 word tests. As shown in the figure, 100 percent of the CID W-22 words are correctly identifed at about 53 dB SPL, while only 92 percent of the PAL PB-50 words are correctly identified at approximately 70 dB SPL (Goetzinger, 1978). In other words, the intelligibility of the W-22 test was higher at comparable intensities than it was for the Rush Hughes recordings (Silverman and Hirsh, 1955). Several reasons for the disparity between these two tests have been given. Basically, the reasons belong to one of two categories, the characteristics of the speaker and the list content (Elkins, 1970).

The articulation of the speaker, Ira Hirsh, on the CID W-22 recordings is superior to that of Rush Hughes, who recorded the Harvard

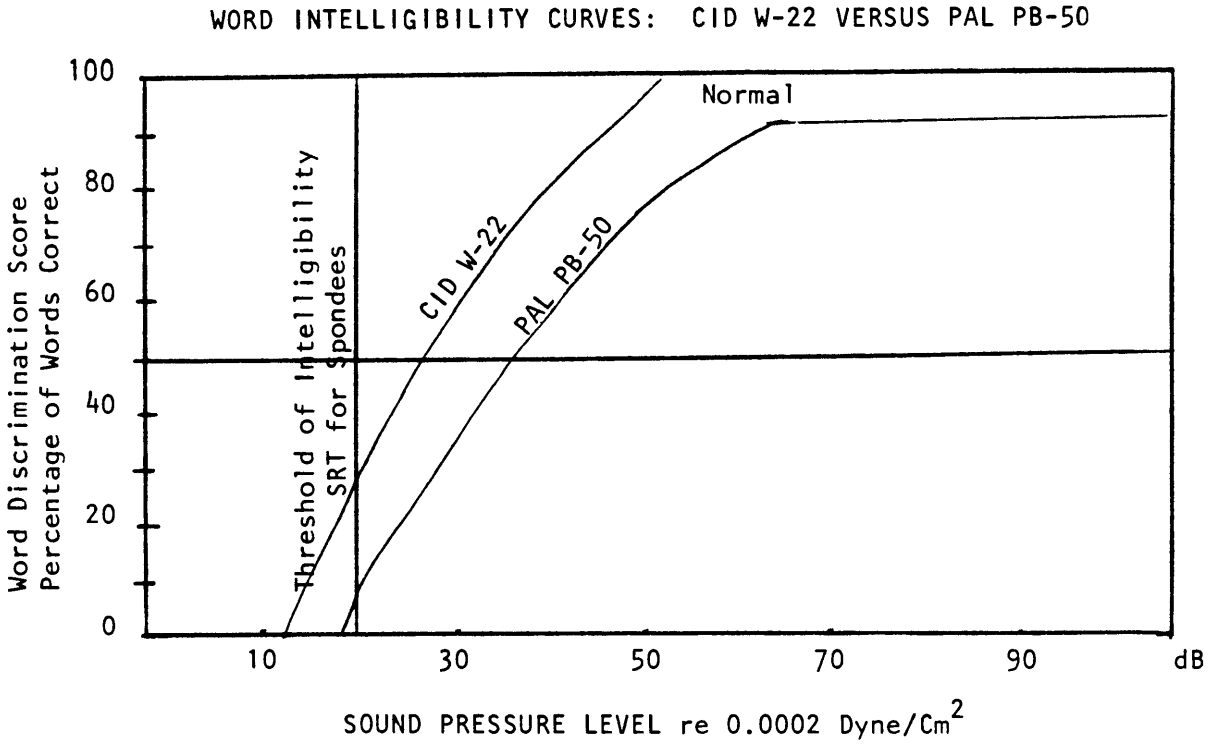

Figure 1. Articulation functions for the CID W-22 and the PAL PB-50 word tests. 100% of the CID W-22 words are correctly identified at approximately 53 dBSPL while the PAL PB-50 words only reach a maximum of 92% intelligibility at approximately 70 dBSPL. (Modified from Goetzinger, 1978).

PAL word lists (Goetzinger, 1978; Davis and Silverman, 1970). Silverman and Hirsh (1955) question that maybe the difference between the recordings is due to the duration of the test words, the Hughes recording containing words much shorter in duration on the average than those of Hirsh.

The W-22 lists contain many familiar words, more than those contained in the PAL PB-50 word lists (Hirsh, et al., 1952; Elkins, 1970; Goetzinger, 1978). Besides being more familiar, they are also fewer in number. The vocabulary for the CID W-22 word lists consists of 200 words as compared to a total of 1000 words in the PAL PB-50 word lists. This restriction in vocabulary reduces the number of alternatives from which to choose, thereby making it an easier test (Hirsh, et al., 1952).

Stevens (1978) believes that the difference between the CID W-22 test and the Rush Hughes recording of the PAL PB-50 word lists was due to the speaker who recorded the material and not due to the restriction in test vocabulary. Due to certain unique vocal characteristics of Rush Hughes, the PB-50 recordings represent a relatively difficult speech discrimination test and consequently never received widespread clinical application (Tillman and Olsen, 1973).

The CID W-22 Auditory Test is probably the most widely used speech discrimination test at present (Elpern, 1960; Ross and Huntington, 1962; Martin and Forbis, 1978; Hodgson, 1980). The CID W-22 test has enjoyed a wide, general usage among various researchers and a great deal of experimental and clinical data has been obtained with its employment (Schultz, 1964). However, there is considerable

clinical evidence that the CID W-22 word lists are too easy to be used in the differential diagnosis of various types of hearing loss (Burke, et al., 1965; Campbell, 1965; Goetzinger, 1978).

Due to the fact that the PAL PB-50 word lists were not adequately phonetically balanced, Lehiste and Peterson (1959) developed a new monosyllabic word test. From a pool of 1263 monosyllabic words of the consonant-nucleus-consonant (CNC) type, they developed 10 lists, each containing 50 words. Later, Peterson and Lehiste (1962) revised the original CNC lists by eliminating some of the unfamiliar words that appeared in the original lists. Elkins (1970) concluded that the CNC word lists are phonetically balanced and relatively uniform regarding word familiarity. Despite this contention, Goetzinger (1978) states that various researchers have been unable to find data substantiating the reliability of the 10 CNC word lists.

Based upon the Lehiste and Peterson CNC word lists, the Northwestern University (NU) auditory test No. 4 was constructed by Tillman, Carhart, and Wilber (Goetzinger, 1978). The NU test No. 4 contains two 50-word lists which research has shown to be interchangeable (Rintelmann, et al., 1974; Goetzinger, 1978). Later, this test was revised and expanded by Tillman and Carhart to include four 50-word lists and was called the NU auditory test No. 6 (Hodgson, 1980, Bess, 1983). Although the NU test No. 6 demonstrates high test-retest reliability and good interlist equivalence (Davis and Silverman, 1970; Rintelmann, et al., 1974; Hodgson, 1980), neither the NU test No. 4 or No. 6 has received extensive clinical use (Tillman and Olsen, 1973; Goetzinger, 1978). A commercial cassette recording of the

NU-6 test is available from Auditec of St. Louis (Hodgon, 1980).

Using the Auditec cassette recordings of the NU-6 and CID W-22 speech discrimination tests, Beattie, et al. (1977) found that both tests were essentially similar in difficulty. Figure 2 shows the articulation function for the CID W-22 versus the NU-6 word tests as recorded by Auditec. As shown in the figure, 95% of the words were correctly identified at 32 dB SL for both tests. The authors concluded that the Auditec recordings of the NU-6 test and the CID W-22 word tests can be used interchangeably.

Rintelmann, et al. (1974) recorded their own version of the NU-6 test and found it to be a little more difficult than the original recorded version by Tillman and Carhart. As noted with Beattie, et al. (1977), the maximum discrimination score is reached at approximately 32 dB SL. In a second experiment, Rintelmann, et al. (1974) found that expanding the NU-6 word test to several forms by word randomization does not change the articulation function, nor does it increase the variability within each of the lists.

Hodgson (1980) cites other authors who found the Auditec recording to be more difficult than the original recording by Tillman and Carhart. These findings, along with those of Rintelmann, et al. (1974) who found their recordings also more difficult than Tillman and Carhart's original recording, support the fact that different recordings of the same test may differ significantly depending on the vocal characteristics of the speaker and the recording conditions (Kreul, et al., 1969).

Intertest List Reliability. Which test to use, even which word

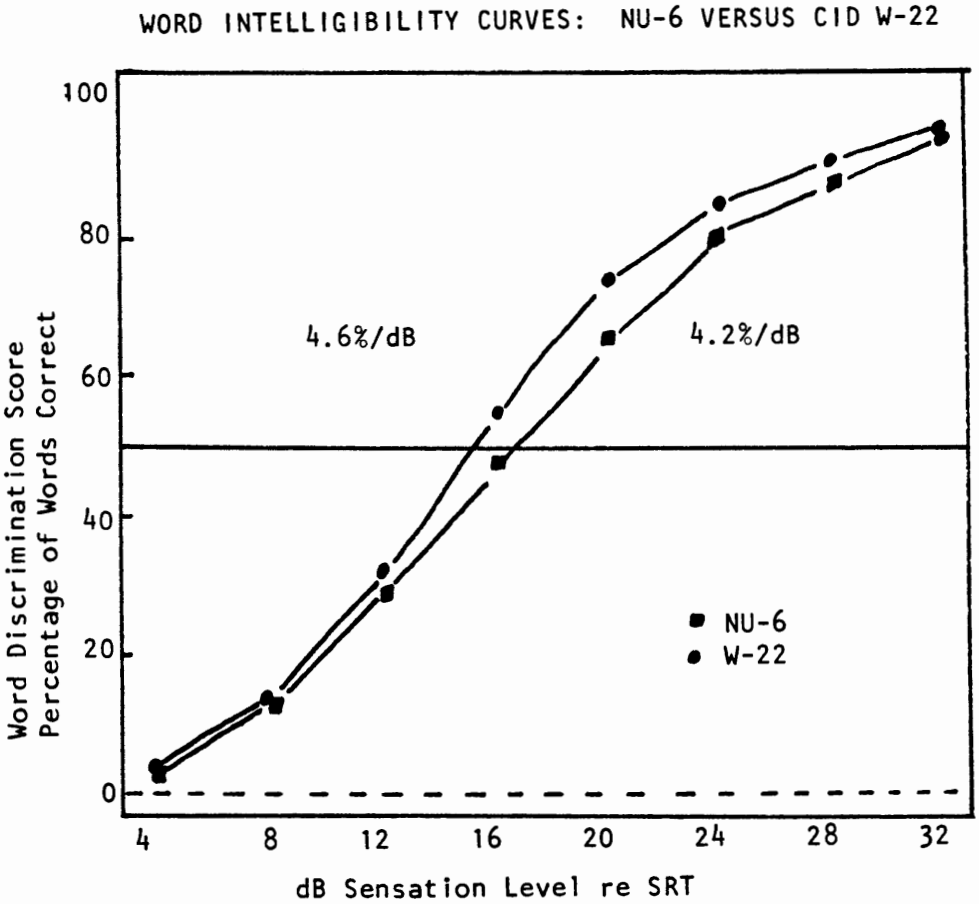

Figure 2. Articulation functions for the NU-6 and the CID W-22 word tests. 95% of the words are correctly identified at 32 dB re SRT for both tests (Modified from Beattie, et al., 1977)

list is a puzzle sometimes. Some of the word lists within each test are not comparable with each other regarding level of difficulty. Burke, et al. (1965) found list 1A, 2A, 5A, and 6A of the PAL PB-50 word lists to be equal in terms of difficulty. Lists 1 and 2, 2 and 3, and 3 and 4 of the CID W-22 word lists were found to be equal in average level of difficulty and therefore considered equivalent for clinical purposes (Elpern, 1960; Ross and Huntington, 1962). Although the NU-6 test has good interlist equivalence, Rintelmann, et al. (1974) found that with respect to degree of difficulty, List 1 was the most difficult, list 2 and 3 were intermediate and essentially equivalent, and list 4 was the least difficult. However, the amount of variability between lists was not greater than the amount of variability found within each list (Rintelmann, et al., 1974).

Full List Versus Half List. There is also some controversy regarding the use of a full 50-word list versus using a half-list containing 25 words. Elpern's (1961) study of the recorded CID W-22 word lists found that the discrimination score for half-lists varies minimally from that obtained with full lists. He stated that the first or second half of a CID W-22 word list may be given instead of the full 50-word list. Burke (1965) also stated that speech discrimination can be measured just as effectively with a 25-word list as with a 50-word list. However, Egan (1948) found it impossible to maintain phonetic composition when using half-lists and argued against their use. Carhart (1965) also argued against using half-lists but for a different reason. Carhart (1965) stated that when a half-list is used only 15 of the 25 words are effective as test items, because the CID W-22 lists

contain too many easy words. He considered this number to be inadequate. Rintelmann et al. (1974) found small mean differences between half-lists and full-lists and stated that half-list testing is warranted when using the NU-6 word lists.

A standard speech discrimination test has not yet been adopted. However, phonetically balanced monosyllabic words have received by far the most widespread clinical application. These monosyllabic word tests have been accepted as a measure of an individual's efficiency in everyday hearing due to their face validity which results from the phonetic balancing (Tillman and Olsen, 1973). It is not yet known whether the phonetic balance contained in the CNC lists and NU test No. 4 and No. 6 versus that used in the PAL PB-50 and CID W-22 word lists will cause significant changes in the clinical interpretation of speech discrimination tests (Goetzinger, 1978). Carhart (1965) found that discrimination scores are not greatly influenced by the change in criteria regarding phonetic balance between the PAL PB-50 word lists and the CNC lists of Lehiste and Peterson. He stated that as long as the test items are meaningful monosyllables with appropriate phonetic distribution, that one 50-word test is relatively equal to another. However, discrimination scores are not stable from one method of presentation to another. For example, the discrimination scores obtained with the Rush Hughes recording of the PB-50 word lists are not comparable with the scores obtained using the Ira Hirsh recording of the CID W-22 word lists (Carhart, 1953b).

Sentences. A third class of speech stimuli used for testing speech discrimination consists of sentences. Jerger, et al. (1968)

stated that the use of single syllable words does not allow for the manipulation of a particular aspect of "ongoing speech," and that is the way speech changes its pattern over time. They stated the necessity for the development of longer samples of speech and consequently developed the Synthetic Sentence Identification Test (SSI). The sentences contained in the SSI are "artifical" meaning that they are not "real" sentences, and "synthetic" meaning that the sequence of words within each sentence follows the rules specified by syntax (Jerger, et al., 1968; Tillman and Olsen, 1973; Goetzinger, 1978). The SSI test is composed of various lists, each containing 10 sentences and utilizes a closed message set, or one in which all possible responses to a particular test stimulus are rigidly specified. The SSI test is seen as an important addition to speech audiometry in that it utilizes a closed message set, scoring is unambiguous, test items consist of multiword sentences rather than single words, and finally, equivalent forms of test stimuli are easily generated (Jerger, et al., 1968).

Egan (1948) stated that the intelligibility of sentences is influenced by a variety of psychological factors such as meaning, context, and rhythm which makes the discrimination score difficult to analyze and interpret. Miller, et al. (1951) stated that words are easier to understand in a sentence than in isolation due to the effect of a restricted vocabulary or contextual cues. Levin (1952) stated that words are preferred as speech stimuli because sentences furnish contextual cues. However, due to the meaningless nature of the synthetic sentences generated in the SSI test, these problems do not

appear to be a factor regarding these particular sentence tests. One other influential variable involved in using sentences as speech stimuli is that of memory. When constructing a group of 100 sentences for the Central Institute for the Deaf, Silverman and Hirsh (1955) limited their sentences to 12 words in length to avoid testing memory span.

There are other sentence tests available such as the Bell Telephone Laboratories sentences and the PAL auditory test No. 8 and No. 12, however, open set sentences are not commonly used in the clinical setting (Davis and Silverman, 1970; Hodgson, 1980). Speech discrimination scores in sentence tests may be higher due to the redundancy supplied by semantic and syntactic cues and not due to actual phoneme identification. Scores might also be reduced by such factors as auditory memory or confusion concerning sentence meaning (Hodgson, 1980). However, monosyllables which are much less redundant or predictable than sentences, require the listener to identify speech sounds independent of contextual and semantic cues, thereby allowing speech discrimination skills to be accurately assessed (Carhart, 1965). It is for these reasons that this examiner choose to use monosyllabic words as the test stimuli.

Method of Presenting Speech Stimuli. One final set of variables encountered in speech discrimination testing concerns the method of presentation of speech materials. One of these variables is the presentation level at which the test is given. The presentation level has an important effect on the discrimination of speech (Berger, 1971). There is still much controversy concering which presentation

level will yield the patient's best speech discrimination score, or PB-Max score. In trying to achieve a patient's PB-Max score, speech discrimination tests are most commonly administered at 25 dB or 40 dB above the patient's SRT. Occasionally other levels are used such as a patient's Most Comfortable Level (MCL) or just below the Loudness Discomfort level or at some other arbitrary level (Carhart, 1965; Berger, 1971).

The level of presentation which will yield a patient's maximum discrimination varies with the test used and its speaker, as well as the individual being tested (Carhart, 1965; Berger, 1971). When using the Hirsh recording of the CID W-22 word lists, PB-Max for normal listeners is usually obtained at a sensation level of 25 dB above SRT. However, when using the Rush Hughes recording of the PB-50 monosyllables, maximum discrimination is not obtained until 40 dB above SRT (Carhart, 1965; Berger, 1971). Maximum discrimination scores for normal listeners appears to occur at 32 dB above SRT when using various recordings of the NU-6 word tests. Rintelmann, et al. (1974) reported PB-Max for normals at 32 dB SL when using his own recording of the NU-6 test. Beattie, et al. (1977) also found PB-Max for normal listeners to be 32 dB SL using the Auditec recording of the NU-6 word test.

Most talkers when using a monitored live-voice presentation obtain results close to those of the CID W-22 recordings. Therefore, many clinicians suggest a presentation level of SRT plus 40 dB when administering the Rush Hughes recordings and a presentation level of SRT plus 25 dB when giving the CID W-22 recordings or when using a monitored live-voice presentation (Carhart, 1965; Berger, 1971; Geffner

and Donovan, 1974). Geffner and Donovan (1974) state that when using CID W-22 word lists, a presentation level of 40 dB above SRT appears to be overamplification and that 25 dB is sufficient. The lack of a standard level for speech discrimination testing makes comparisons between clinics and research reports more difficult (Berger, 1971).

Another factor concerning the method of presenting speech stimuli which has an influence on test results is the use of a carrier sentence. The use of a carrier sentence is desirable for several reasons. It prepares the listener for the presentation of the test stimulus and therefore reduces variability in the speech discrimination score due to inattention and distraction (Egan, 1948). The use of a carrier sentence also allows the speaker to modulate the level of his voice and keep it even from word to word. Monitoring the carrier sentence "You will write . . ." or "You will say . . ." rather than the test stimulus preserves the relative intensity of the speech sounds in a natural manner (Egan, 1948; Miller, et al., 1951).

The use of the carrier phrase "You will say . . ." may actually contribute to the intelligibility of the test word. Lynn and Brotman (1981) found that the carrier phrase "You will say . . ." contains acoustic information that can help the listener identify place of articulation for initial voiceless stop consonants of some of the CID W-22 test words. They also found that when the carrier phrase was removed, it produced a more difficult word discrimination test. Therefore, Lynn and Brotman (1981) stated that test scores with carrier phrases cannot be compared to those administered without carrier phrases due to the differences in test difficulty. They also stated

that different carrier phrases would most likely produce different results.

Summary of Speech Audiometry

Speech discrimination testing has proven to be a highly useful diagnostic tool in the audiological evaluation. However, several problems exist regarding the present state of the art of speech discrimination testing. Various problems are encountered due to the number of different materials used as speech stimuli and the method of their presentation. Traditionally, speech materials have consisted of single words, either bisyllablic spondee words for assessing speech reception threshold, or monosyllabic PB word lists for assessing speech discrimination ability. Administration of these speech materials has typically been through recordings such as those of Rush Hughes for the PAL PB-50 word lists, the CID W-22 word lists recorded by Ira Hirsh, the Auditec recording of the NU-6 word tests, or by monitored live-voice. None of these methods of presentation are considered totally satisfactory due to the numerous amount of variables encountered with their administration.

At present, the existing tests for speech discrimination and their method of presentation lack standardization. What is needed is a single test instrument and method of administration that will combine the best qualities of the previous tests and their methods of presentation, thereby eliminating the variables that have plagued traditional speech discrimination testing. Utilizing synthetic speech in testing speech discrimination is a possible answer to the problems encountered with traditional speech discrimination tests. Synthetic

speech could eliminate the variablity encountered with monitored live-voice as well as that involved with recorded speech materials. Synthetic speech could provide uniform, standardized speech stimulus materials, easily generated, and comparable between clinics and research laboratories.

SYNTHETIC SPEECH

Speech synthesis is the method by which intelligible, human-like speech is generated from a machine (Lerner, 1982; Canning, 1983). Using a computer to generate speech gives us a new freedom and greater flexibility. It allows us to create new words and sentences without the use of recorded speech materials (Morgan, 1984). In other words, speech can be created where none existed previously.

Man has been interested in making machines capable of reproducing the human voice for the past two centuries. As far back as 1779, innovators such as Kratzenstein and Von Kempelen constructed talking machines which used vibrating reeds simulating the vocal cords. These reeds in turn were connected to a set of acoustic resonators representing the human vocal tract. Although the speech produced was poor in quality, certain discrete sounds were able to be produced from these machines (Ainsworth, 1976; Morgan, 1984).

It was not until the advent of electronics that more successful talking machines were produced. One of the first electrical synthesizers was Homer Dudley's VODER (Voice Operation DEmonstratoR). The VODER was played like a keyboard instrument by a trained operator and produced intelligible speech resembling the human voice (Ainsworth, 1976; Morgan, 1984).

Today, electronic and automatic devices, some of which are contained on a single integrated circuit, are capable of producing speech. The production of intelligible speech by a computer is no longer a novel idea and the theories involved in speech synthesis have not changed a great deal from those used in designing the VODOR (Lerner, 1982; Morgan, 1984).

There are three basic methods of producing synthetic speech (Canning, 1983). The first method involves waveform digitization and compression in which an actual speaker's voice is digitized. It is a process whereby human analog speech is sampled and converted into a sequence of binary numbers (analog-to-digital, or AD, conversion) (Kamm, et al., 1980). In actuality, a speaker's voice is tape recorded and the waveform of that voice is then digitized. Therefore, this technique reproduces a speaker's voice rather than generating actual synthetic speech. Digitizing a speaker's voice, storing it and replaying it is not that much different from a tape recording (Canning, 1983; Darrow, 1983). However, digitizing speech allows it to be compressed and manipulated in ways that conventional tape recorded or analog speech cannot (Canning, 1983). For example, digital speech can be speeded up without affecting the frequency area of the speech (Rahko, et al., 1979). In other words, speech that has been digitized can be manipulated in various ways without causing the speech sample to be distored. Digital speech is very life-like and highly intelligible but the process requires large amounts of storage (Nakatsui and Hermelstein, 1982; Canning, 1983; Darrow, 1983).

The next two methods of producing synthetic speech are based on

the properties of the human vocal tract. These methods include linear predictive coding (LPC) and formant synthesis. These techniques are based on the model that the power of the speech signal is concentrated in several frequency bands which correspond to the resonances of the human vocal tract. By specifying the main frequencies and their respective intensities a good approximation of the speech signal is obtained (Lerner, 1982).

In linear predictive coding, speech is represented by mathematical models (linear prediction coefficients) which are based on actual spoken words. Linear prediction is based on the fact that since speech is highly redundant, a current sample of speech can be synthesized by predictions from a weighted linear combination of previous samples (Damper, 1982). These mathematical models include information regarding the frequency, intensity, duration, and voiced and un-voiced aspects of speech (Damper, 1982; Canning, 1983). Linear predictive coding techniques have been reported to produce good quality speech (Canning, 1983).

The third and final technique of synthesizing speech is called formant synthesis. It is a technique which models the resonances or formant frequencies of speech by input parameters which directly define the frequencies, bandwidths, and amplitudes of the main frequency bands or formants. These input parameters are then used to drive digital filters (Damper, 1982; Lerner, 1982). Formant synthesis is a process in which speech is broken down into formants which are pieced together again to form words (Canning, 1983). At least three formants are needed to produce acceptable speech with the addition of resonances to

model nasal and fricative sounds (Damper, 1982). Formant synthesis is used to create speech where no speech existed previously, as in speech processed from written text. Formant synthesis techniques provide the possibility of an unlimited vocabulary, however, the speech is intelligible but very mechanical in quality (Allen, 1981; Damper, 1982; Canning, 1983; Darrow, 1983). In fact, it produces the poorest sounding speech of all three methods (Canning, 1983).

Although both linear predictive coding and formant synthesis implement vocal tract models, a major difference between the two techniques is the way in which they use the model of the vocal tract to describe the speech signal. Linear predictive coding uses LPC coefficients to describe the speech sample while formant synthesis involves formant frequencies and amplitudes (Damper, 1982). Lerner (1982) states that the main difference between linear predictive coding and formant synthesis lies in the treatment of the digital filter. Linear predictive coding uses a complex digital filter which processes several different frequencies in one step. On the other hand, formant synthesis uses a set of individual frequency filters through which the signal is passed in a sequence or in parallel (simultaneously) and then added (Lerner, 1982; Morgan, 1984). The difference between linear predictive coding and formant synthesis versus speech that has been digitized is that the former techniques process speech on a phoneme by phoneme or individual speech sound basis whereas digitized speech is processed on a word by word basis.

Recently, attention in speech synthesis research has shifted from designing the actual speech synthesizers to the generation of the rules

or parameters which control them (Ainsworth, 1976). This well-known approach to speech synthesis is called synthesis-by-rule (Leng, et al., 1981). It is a method of generating speech through the application of a set of rules to an input string of phonetic elements (Ainsworth and Millar, 1976; Allen, 1981; Damper, 1982). These segments may consist of phonemes or syllables (Liberman, et al., 1959). Phonemes are discrete elements which represent the minimal sounds from which all words are created for a given language (Ainsworth, 1976; Allen, 1981; Damper, 1982). Liberman, et al. (1959) states that by writing the rules for speech synthesis in terms of phonemes rather than syllables, the number of rules needed is considerably reduced.

In using a synthesis-by-rule approach, a set of rules have been drawn up which allow for the automatic calculation of the parameters which are necessary to synthesize a given string of phonemes (Ainsworth, 1976). Since a description of each phoneme is stored, any word can theoretically be produced. The stored data regarding phonemic descriptions can be in LPC coefficients as in linear predictive coding or in formant frequencies and amplitudes as used in formant synthesis techniques (Damper, 1982). In other words, a synthesis-by-rule approach is a process of generating synthetic speech from text using either linear predictive coding or formant synthesis techniques.

Importance of Speech Synthesis

Advances in computer technology have dramatically increased the number of potential applications for synthetic speech (Morgan, 1984). Speech technology is comprised of speech input and speech output technologies (Darrow, 1983). Speech input or speech recognition

technologies are concerned with the ability of computers to understand human speech. Speech output or speech synthesis technologies, in which synthesized speech emanates from computers or products is the technology with which this study is mainly concerned. However, both technologies apply to man-machine communication, and the advances which occur in either technology will have an effect on the future benefits derived for society from speech synthesis systems.

Due to the advances in integrated circuit technology, there are now several single-chip speech synthesizers available which offer the user the advantages of synthesizing speech using a reliable, low cost, portable speech synthesizer (Damper, 1982; Morgan, 1984). Although the speech synthesized by a computer does not yet sound fully natural, the quality is improving rapidly (Lerner, 1982; Teja, 1982). It is thought that in the next few years, further advances in computer technology will create an abundance of complex speech systems which will have an affect on almost every aspect of our daily lives (Pisoni, 1982).

Speech synthesis systems are already employed in many diverse fields. Currently, the most common application regarding speech synthesis systems is in business and industry (Damper, 1982; Prado, 1983). Here, speech synthesis systems, including speech synthesis and recognition, are used in data processing and telephone-answering equipment (Elovitz, et al., 1976; Damper, 1982; Pisoni, 1982).

Speech synthesizers are also being implemented in many products that we encounter in our daily lives. It is not uncommon for household appliances, such as toasters, microwave ovens, and alarm clocks to talk. Voice is being given to such things as dashboards in cars,

elevators, vending machines, cash registers, and even to childrens' toys and games (Pisoni, 1982; Lerner, 1982; Darrow, 1983; Morgan, 1984). Robots are even being given the ability to talk. Leng, et al. (1981) describe a robot which has a vocabulary of 125 words. We are beginning to see speech synthesis being implemented in automatic travel reservation and credit verification systems as well as in automatic bank teller operations (Kitawaki, et al., 1981; Lerner, 1982; Pisoni, 1982). Speech synthesis is definitely revolutionizing the world in which we live.

Speech synthesizers are beginning to be used widely in educational areas. Computer-assisted instruction (CAI) using synthetic speech is being used to help beginning readers (Pisoni, 1980; Cumming and McCorriston, 1981). Certain books have synthetic speech capabilities. The "Magic Wand Speaking Reader" manufactured by Texas Instruments allows bar codes to be converted into synthetic speech as the child reads the text, thus allowing him to associate the written words with the sounds they represent (Darrow, 1983). Various educational games employing speech synthesizers are available. One such game is the Talking Learning Computer from Tiger Electronic Toys which asks children questions regarding mathematics, spelling, and reading, and then identifies their typed response as right or wrong (Lerner, 1982).

An area in which synthetic speech seems to have many valuable applications is in assisting and providing aids for the handicapped. One such application involves the use of a manually operated speech synthesizer which provides speech for the mute or speech impaired user

(Levitt, 1980; Rahko, et al., 1980; Damper, 1982; Darrow, 1983). Speech synthesis also provides a way to present visual information to the blind (Damper, 1982). Several products such as the talking typewriter, calculator, and computer terminal allow the blind to compete in the everyday work world. The Talking Typewriter by IBM allows the blind typist to prepare and proof-read reports. The "talking computer terminal", also by IBM, enables the blind to be employed as computer programmers (Damper, 1982). A major application of speech synthesis concerning the blind is the development of a reading machine for the blind. One such machine, produced by Kurzweil Computer Products in 1979, optically scans printed text materials and then converts them to speech output (Levitt, 1980; Damper, 1982, Lerner, 1982). Although the speech quality of this reading machine is not very good, the user can adjust to it quite easily after an initial learning period (Damper, 1982; Lerner, 1982).

Speech recognition techniques are being used to help the physically disabled in their home environment. Such products as a voice-operated wheelchair and environmental control unit (ECU) are being investigated. However, due to the state of the art regarding speech recognition technology, speech-recognition aids are not available commercially (Damper, 1982).

Computerized speech technology has many applications in the fields of speech pathology and aural rehabilitation (Levitt, 1980; Rahko, et al., 1980). Since only about 25 percent of spoken syllables can be determined through speech reading techniques, the use of computerized technology to provide supplemental speech information in

the form of visual or tactile cues has proven very useful (Levitt, 1980; Damper, 1982). An ideal speech aid for the deaf would translate the speech signal of others into text for visual display. Although this is not possible at the present time due to the limitations of present speech recognition techniques, modifications of this principle are possible with current technology. Aids based on this principle find numerous applications both in communication aids for the deaf as well as in speech training for deaf and hearing impaired children (Damper, 1982). Computerized speech technology can be used to recognize and extract important components of the speech signal and present such information automatically in the form of visual and tactile cues as a supplement to the speech reading or speech training process. This information can be derived from the child's own speech or from the speech of a normal child (Levitt, 1980; Damper, 1982). Information concerning fundamental frequency, nasality, intensity, and speech spectrum can be displayed (Levitt, 1982). There is one computerized speech-training aid available which provides a visual schematic of the speaker's vocal tract (Levitt, 1980; Damper, 1982). These visual and tactile computer displays represent target traces which the deaf or hearing impaired child tries to match or reproduce (Damper, 1982). These computerized displays may also be used in examining important phonetic characteristics of a child's own speech (Levitt, 1980; Rahko, et al., 1980).

An area in which synthetic speech has potential value is that in which the natural speech signal is processed and enhanced in such a way as to make it more intelligible to the hearing impaired. The speech

signal may be enhanced by such signal processing techniques as frequency lowering, extraction of formants, frequency-dependent compression amplification, and background noise reduction. While there have been some improvements in this area, no major breakthroughs have occurred (Levitt, 1980).

Other devices, employing speech synthesizers are being used to assist the deaf or hearing impaired. One such device, called the speech-text synchronizer, allows the deaf to appreciate the rhythmic nature of speech by viewing text which is synchronized with the syllables as they are spoken. The speech information can be presented in the form of tactile vibration, as real-time spectrograms of speech, or as sound using a digital speech synthesizer depending upon the preference and degree of hearing loss of the user (Sargent, 1982). A discussion or speaking machine, which employs a speech synthesizer and keyboard with a remote terminal, is currently under development for those with reduced hearing or difficult speech (Rahko, et al., 1980).

Speech synthesis techniques may also be used to simulate the speech of hearing impaired children in order to localize the various speech problems and therefore help provide a plan of remediation for producing the needed improvements in the child's speech. Computerized methods also allow for objective evaluation of data regarding a child's speech production abilities. Subsequently, this information may be used in planning speech therapy and measuring a child's progress (Levitt, 1980).

Another area in which computer technology is beginning to be implemented is in the field of audiology. According to Levitt (1980),

the potential applications of computer-assisted testing in audiology was recognized over ten years ago. However, few clinics were able to afford the expensive computer systems at that time. Recently, the advances in microprocessor technology have made computer technology more affordable and practical for use in auditory testing (Levitt, 1980). Computers are beginning to play an ever-increasing role in audiology, and computer-assisted testing is beginning to be implemented in the clinical setting. Many clinics are currently investigating the use of computers to assist in Bekesy audiometry, auditory brainstem response (ABR) audiometry, and in impedance and acoustic reflex testing (Kamm, et al., 1980; Levitt, 1980).

Computerized methods allow for a greater degree of objectivity in various aspects of speech discrimination testing. Some of these aspects include the generation of the speech stimuli as in speech synthesis, control of the tests, response analysis, corrective feedback, and automatic interpretation of the results (Kamm, et al., 1980; Levitt, 1980). Computer-aided procedures are available which provide assistance in the recording as well as in the delivery of speech materials for audiological research (Kamm, et al., 1980).

Rahko, et al. (1979) and Damper (1982) have indicated that computer-generated or synthetic speech is likely to prove useful in the audiological evaluation and the diagnosis of hearing impairment. One of the major advantages of synthetic speech over that of conventional pre-recorded speech materials is its greater flexibility (Kamm, et al., 1980; Damper, 1982). Synthetic speech can be manipulated in a variety of ways such as by compression, speeding, slowing, mixing, editing,

timing, measuring and filtering (Rahko, et al., 1979; Kamm, et al., 1980). Due to its versatility, synthetic speech has the potential for being used in several different audiological testing applications. For example, synthetic speech can be speeded up without affecting the frequency area of the speech. It is the only way of producing speech which can be thus manipulated without causing severe transient, consonant, and vowel pitch difficulties (Rahko, et al., 1979). Therefore, it may prove useful as a diagnostic test for central auditory problems due to the decreased redundancy in the speech signal. Synthetic speech also allows for the controlled variation of perceptually significant cues, such as formant transitions, thereby permitting the investigation of their effects on hearing impairment (Godfrey and Millay, 1980; Ginzel, et al., 1982).

Another advantage of synthetic speech over recorded materials is it greater degree of objectivity (Kamm, et al., 1980). Synthetic speech avoids the variability that is inherent in naturally spoken or recorded test materials due to the differences of individual speaker's voices (Godfrey and Millay, 1980). The use of computer-generated speech allows for a more objective speech stimulus to be created and therefore, may be a possible solution to the variables encountered with the use of monitored live-voice or recorded speech materials.

Synthetic speech may also prove useful in hearing aid evaluations. Godfrey and Millay (1980) state that they hope to develop a hearing aid evaluation procedure using standardized synthetic speech materials. According to Levitt (1980), a computerized system could be set up to systematically check the calibration of various hearing aid settings.

The importance of synthetic speech has been well established. It is beginning to be seen and used in products which we encounter in our everyday lives (Pisoni, 1982; Darrow, 1983; Morgan, 1984). Due to the multitude of its various applications, speech technology, in particular speech output or speech synthesis is helping disabled people, including the non-vocal, the blind and visually impaired, the physically disabled, and the deaf to become productive members of our society (Damper, 1982; Williams, 1982). Speech synthesis is making it possible for the handicapped to receive an education and gain employment, and thus contribute to their community (Williams, 1982). Computerized speech technology is beginning to be used in the fields of speech pathology, audiology and aural rehabilitation (Levitt, 1980). Computers are being used in these fields to perform computer-assisted testing, for speech anlaysis and synthesis, as supplemental aids to speech reading and speech training, and to help plan and evaluate various speech training programs (Kamm, et al., 1980; Levitt, 1980). Of particular interest to this study is the application of synthetic speech in audiology.

Studies of Intelligibility Regarding Synthetic Speech

From the above applications, it can be seen that computerized technology, including the use of synthetic speech materials is beginning to be implemented in the field of Audiology. Therefore, it does not seem untimely to look at several of the studies dealing with the intelligibility of synthetic speech with the hope that it is possible to use synthetic speech in the audiological investigation of hearing impairment.

According to Nye and Gaitenby (1973) there are three major variables involved when testing synthetic speech. These variables include the rules of synthesis (phonemes, allophones, prosodic features), the synthesizer used to produce the synthetic speech stimuli, and the test procedure which is used to evaluate intelligibility. Therefore, comparison between studies is risky due to the synthesizer used, the test materials, and their administration. However, it is necessary to look at the results of various intelligibility studies in order to evaluate the present effectiveness of synthetic speech.

The production of synthetic speech from written text has been studied by several researchers and organizations for various applications. Since the major applications of speech synthesis systems have been in business and industry (Damper, 1982; Prado, 1980), several of the studies of intelligibility regarding synthetic speech have been with speech synthesizers employed in these areas (Ainsworth, 1971; Ainsworth and Millar, 1976).

Ainsworth (1971), used an expensive formant synthesizer similar to the one developed for use by Haskins Laboratories to synthesize spoken letters of the alphabet as the test stimuli instead of phonetic codes. He found the mean intelligibility of the sounds to be 83% for normal hearing subjects. None of the errors were due to vowel recognition. However, weak fricative sounds were not well produced which lead to such confusions as F with L and T with G.

In a later study, Ainsworth and Millar (1976) investigated the effects of synthesizing speech by rule from phonetic data which

includes rules for generating consonants dependent upon their surrounding context, in this case the vowel following the consonant. When taking this contextual factor into account, the intelligibility of the stop consonants /b, d, g, p, t, k/ in isolated consonant-vowel (cv) syllables rose from 68% to 92%.

A major portion of the research on the intelligibility of synthetic speech has been concerned with the intelligibility of individual words or words embedded in single sentences (Jenkins and Franklin, 1982). A great deal of research concerning the intelligibility of synthetic speech has been done by the Haskins Laboratories using a general purpose speech synthesis system. This system for generating synthetic speech is composed of an expensive computer system coupled with a parallel formant resonance synthesizer (Nye and Gaitenby, 1973). Most of the research work done at Haskins Laboratories has been concerned with the generation of synthetic speech for psycholinguistic experiments and for use in reading machines for the blind (Mattingly, 1968; Nye and Gaitenby, 1974; Nye, et al., 1975). The tests employed to assess the intelligibility of synthetic speech were chosen to pinpoint errors in synthetic speech and not to assess hearing. Also, the presentation levels varied between studies, and sometimes were not reported at all.

Nye and Gaitenby (1973) used the Modified Rhyme Test (MRT) and synthetic speech produced by the Haskins Laboratoies' parallel formant resonance synthesizer. They found that the overall intelligibility score for monosyllable words in normal subjects was 92.4% for synthetic speech versus 97.3% for natural speech when administered at 80 dB SPL.

The overall consonantal error rate for both initial and final positions combined was 7.6% in synthetic speech and 2.7% in natural speech. Figure 3 shows the percentage of error for each phoneme that was confused in more than 2% of its appearances. The phonemes /v, ð, θ, p, b/ produced confusions both initially and finally in synthetic speech. The synthesized phonemes that were the least intelligible were initial /v/ and final /r/. The classes of phonemes that needed the most improvement were the labials, labiodentals, and dentals in both initial and final positions. The manner synthesized the least well was frication. The four synthesized phonemes /t, g, f, s/ were clearly intelligible in either position. Figure 4 shows the highly intelligible synthesized phonemes. As a group, the alveolar phonemes were perceived the best, while the labials and labiodentals were the least intelligible due to the formant energy relative to the vowels.

Nye and Gaitenby (1973) also found that subjects showed significant evidence of learning in successive exposures to synthetic speech. They also noted several limitations of the MRT when testing the intelligibility of synthetic speech such as the uneven representation of initial and final consonants and an incomplete representation of vowel environments. Also, because the alternative word choices did not necessarily contain acoustico-phonetically similar consonants to the consonant in the stimulus test word, the substitutions may not be valid comparisons. Therefore, only the phonemes that produced the error and not the phonemes with which they were confused could be used for modifying the rules for speech synthesis. In other words, the MRT was not good for identifying the

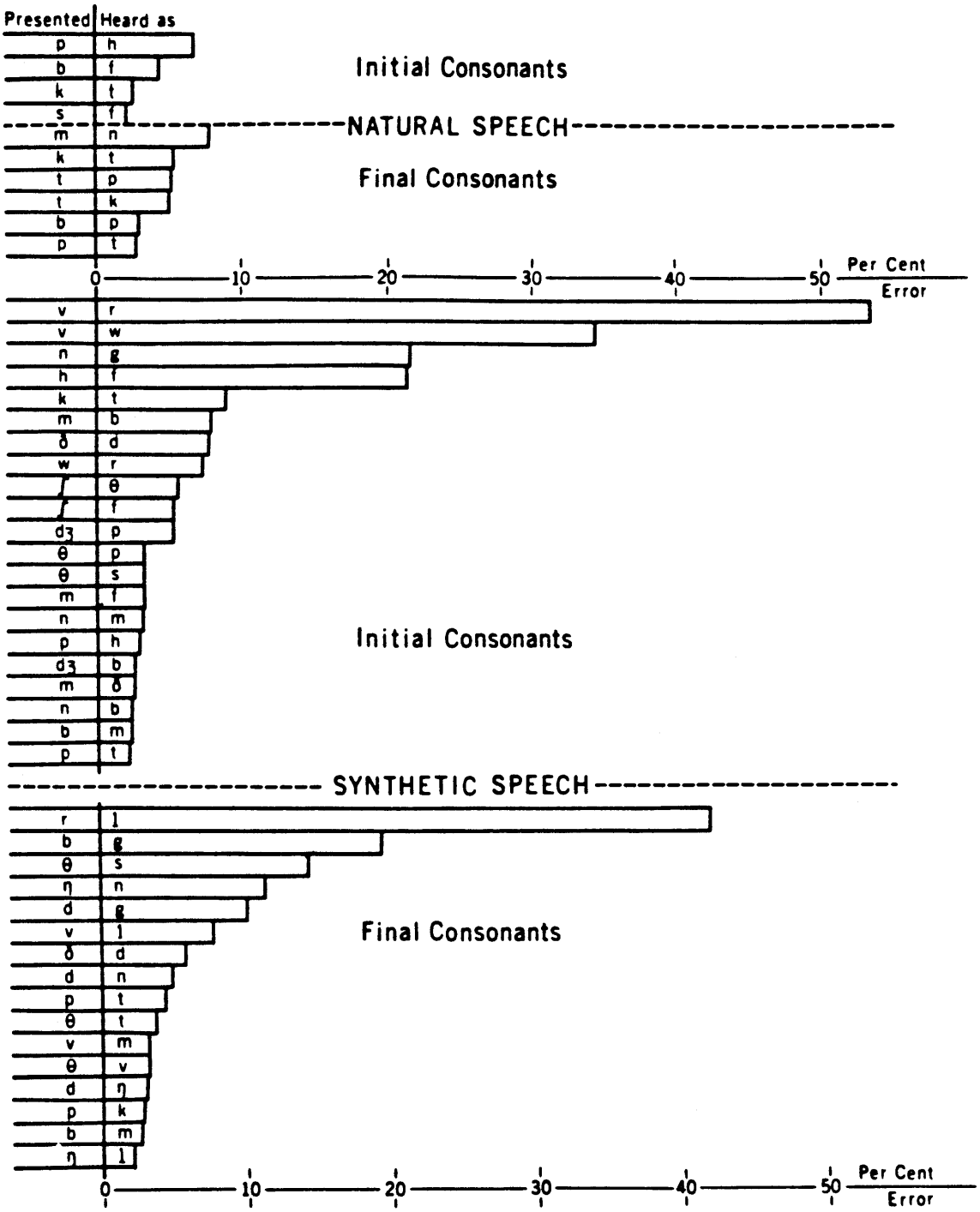

Figure 3. Phonemes ranked by percentage of error (Nye and Gaitenby, 1973).

	LABIAL	LAB-DENT	DENTAL	ALVEOLAR	PALATAL	VELAR	GLOTTAL
STOP Voiceless				t		k^2	
STOP Voiced				d^1		g	
FRICATIVE Voiceless		f		s	$tʃ^{*2}$		
FRICATIVE Voiced				z^{*2}	$dʒ^2$		
NASAL	m^2			n^2			
LIQUID				r^1			

LEGEND

Phoneme presented in final position only: *

Phoneme good in initial position only: 1

Phoneme good in final position only: 2

(/ʒ/ and /y/ did not appear in the test.)

Figure 4. Highly intelligible synthesized phonemes (Nye and Gaitenby, 1973).

confusions that occur among poorly perceived phonemes in synthetic speech.

In a second study, Nye and Gaitenby (1974) investigated the intelligibility of synthetic speech (phonemes in context) using the Syntactically Normal Sentence Test (SNST) which contains 200 syntactically normal but semantically anomalous sentences arranged into four series of 50 sentences each. The sentences contained in this test are also referred to as the Haskins Anamalous Sentences (Morgan, 1984). The test sentences were synthesized at a speaking rate of 130 words per minute again using Haskins Laboratories' parallel formant resonance synthesizer. The test sentences were administered at 80 dB SPL and subjects were asked to recall the four key words from each sentence directly after their presentation. The overall average error rate for synthetic speech jumped to 22% compared with that of 7.6% previously obtained with the Modified Rhyme Test, while the overall average error rate only rose from 2.7% for the MRT to 5% for the SNST using natural speech. These results indicate that it is more difficult to recall synthetic speech stimuli in a nonredundant context such as the SNST and that this degree of difficulty is not predictable from the intelligibility results of synthetic speech in the single-word closed-response format of the MRT (Jenkins and Franklin, 1982). The least intelligible phonemes for synthetic speech in the SNST experimental recall test were /θ, tʃ , ʃ, t/ in initial position and /p, g, ŋ / in final position. As found in their previous study using the MRT, the overall intelligibility of final consonants was better than that for initial consonants in synthetic speech. Various limitations

of the SNST such as the "coarticulation" effects at word boundaries (ie., sporadic interword effects) and the meaningless nature of the sentences which may have led the listeners astray by causing them to misinterpret a word, are thought to have contributed to the high overall error rate. Due to the "coarticulation" effects, the intelligibility of initial and final phonemes may be more significantly lower in context than in isolation. Memory is also thought to have been an influential variable in the test results, as well as the fact that extra attention is needed for interpreting synthetic speech. Nye and Gaitenby (1974) also noted the effects of training, in that the new subjects made many more phonetic errors than did the old subjects. However, performance differences between old and new subjects indicates that while inexperienced listeners misinterpret certain synthesized phonemes, the more experienced listeners can adapt to some of the phonetic deficiencies of these same sounds.

In a subsequent study, Nye, Ingemann, and Donald (1975) looked at listeners' ability to comprehend synthetic speech passages in order to see if redundancy compensates for losses in phonetic intelligibility. This study involved extended listening to find out how well listeners could understand and obtain information from synthetic speech. College text materials were synthesized using three different synthesis-by-rule programs and two different synthesizers, the OVE-III serial formant synthesizer and the older Haskins Laboratories' parallel formant synthesizer. The material was synthesized at 133 to 154 words per minute, while the natural speech was recorded at a speaking rate of 170 words per minute. After listening to the passages, listeners were

timed as they answered questionnaires designed to assess their comprehension. The results suggested that listeners required 23% more time to understand the synthetic speech passages versus the natural speech passages. However, the differences in favor of natural speech may have been partially due to the faster speaking rate.

Ingemann (1978) has done several research studies using various word and sentence tests with speech synthesized by the newer OVE III serial formant synthesizer developed at Haskins Laboratories. The program used to synthesize speech with the OVE III synthesizer is called FOVE. It is an elaborate program designed to allow the user to test various acoustic-phonetic rules for synthesizing speech. Ingemann (1978) has performed several different tests on listeners unfamiliar with synthetic speech in order to determine the intelligibility of speech presently synthesized by this system as well as to pinpoint the specific sounds which needed improvement. The overall results of these tests are given in Table I.

In the first study, Ingemann (1978) used a set of meaningful but frequently odd sentences called the Carnegie-Mellon University (CMU) sentences. Listeners unfamiliar with synthetic speech wrote down the sentence after hearing it twice. The results indicated that listeners were able to get 84% of the words correct and 91% of the phonemes correct, with vowels being slightly higher in intelligibility.

In a second study, Ingemann (1978) obtained somewhat lower intelligibility scores for phonemes heard in isolated words only once. Using the Mitchell test (Mitchell, 1974), which contains four lists of 50 monosyllabic words each and designed to test 22 initial consonants,

TABLE I

PERCENT CORRECT ON VARIOUS INTELLIGIBILITY TESTS FOR SPEECH SYNTHESIZED BY THE FOVE PROGRAM USING THE OVE III SYNTHESIZER AND 1977 RULES (Modified from Ingemann, 1978)

	1977 Rules
CMU Sentences	
Words	84%
Phonemes	91%
Consonants	90%
Vowels	93%
Mitchell Lists	
22 Initial Consonants	79%
13 Final Consonants	84%
15 Vowels and Diphthongs	99%
Total	86%
SPD Sentences	
Words	85%
Phonemes	89%
Syntactically Normal Nonsense Sentences	
Words	77,82%
Phonemes	89,91%

13 final consonants, and 15 vowels and diphthongs, phonemes were found to be only 86% correct as compared to 91% in sentences.

In a following study, Ingemann (1978) used a subset of phonetically diverse (SPD) sentences which were designed to contain sounds similar to those in the Mitchell test but were poorly represented in the CMU sentence test. The test consisted of 13 sentences and the scores were similar to those obtained using the CMU sentences with a slightly higher word score and a slightly lower phoneme score.

Ingemann (1978) performed a further study of phoneme and word intelligibility in nonsense sentences using 50 sentences of the SNST developed by Nye and Gaitenby (1974). Two sets of scores were obtained for these sentences as heard only once by two different groups of listeners at their individual comfort level. The first scores were obtained when a group of twelve listeners unfamiliar with synthetic speech heard the 50 sentences of the SNST at the beginning of the test session. The second set of scores were for a different group of twelve listeners not regularly exposed to synthetic speech who heard the same sentences but the sentences were preceeded by ten similar sentences used to test natural versus durational rules (Ingemann, 1979). In these two series of tests, word intelligibility was found to be 77% and 82% while phoneme intelligibility was 89% and 91% respectively. The increase in the second score, even though obtained for different listeners, revealed an obvious learning effect which resulted from the second set of subjects hearing the ten durational rule sentences before the 50 sentence test.

In comparing the word intelligibility of the various sentence tests in Table I, it was noted that the intelligibility of words was lower in the nonsense sentence test (77% and 82%) than in the meaningful sentences test (84% for CMU sentences and 85% for SPD sentences). Phoneme intelligibility was found to be approximately equal in all three sets of sentences (89% to 91% for nonsense sentences, 91% for CMU sentences, and 89% for SPD sentences). However, in isolated words, such as those contained in the Mitchell test, phoneme intelligibility was only 86%.

Ingemann (1979) reported in a subsequent article additional data concerning a learning effect taken from the two scores in the SNST experiment. The words correct in the first 25 sentences were compared with those correct in the last 25 sentences for both test conditions. The results (see Table II) showed that the listeners definitely understood more of the words in the second half (83% and 85%) than they did on the first half (71% and 78%) for the two experimental conditions. The second half of the sentences had 12% and 7% higher word intelligibility for the two experimental test situations. Since the difficulty of the sentences is assumed to be equal between the first and second half of the test, it seems evident that a learning effect did occur. It is interesting to note that this learning effect was not as noticeable when the ten durationally modified sentences preceded the 50-sentence test, which is an indication that a significant amount of learning had already taken place and therefore, not as much occurred between the first and second half of the list.

Ingemann (1979) also examined the effects that more natural

TABLE II

PERCENTAGE OF WORDS CORRECT IN SYNTACTICALLY NORMAL NONSENSE SENTENCES FROM THE SNST AS SYNTHESIZED USING THE FOVE PROGRAM WITH 1977 RULES AND THE OVE III SYNTHESIZER
(Modified from Ingemann, 1979)

	Sentences 1-25	Sentences 26-50	All 50 Sentences
50 sentence test followed by 10 durational sentences	71%	83%	77%
10 durational sentences followed by 50 sentence test	78%	85%	82%

durational rules had on the intelligibility of synthetic speech. The ten sentences were modified so that their segment durations more closely approximated those of real speech. Ingemann (1979) found that although more natural durational rules improved speech scores initially, once learning had taken place the natural durations produced only negligible improvement. She concluded that unnatural durations may cause comprehension problems in the beginning, but that a listener can adapt quickly without special training. On the other hand, she thought that it was important to improve durational specifications for initial acceptance and for listeners who were not expected to listen extensively to synthetic speech.

According to the intelligibility results of the tests mentioned above, Ingemann (1978) states that the FOVE program utilizing the OVE III synthesizer is capable of producing fairly intelligible speech, especially after listeners have been given a short adaptation period. The OVE III synthesizer produces remarkably good vowels, but nasals and fricatives do not closely approximate natural speech sounds. The voiced fricatives are very poorly produced (Ingemann, 1978).

A more recent analysis of synthetic speech has been conducted by Pisoni and Hunnicutt (1980) in the evaluation of the MITalk text-to-speech system. MITalk employs a phoneme-based synthesis-by-rule system which controls a parallel formant synthesizer designed by Klatt (1980). Pisoni and Hunnicutt (1980) initially evaluated MITalk in a fashion closely resembling that of the Nye and Gaitenby (1973, 1974) studies previously described. Pisoni and Hunnicutt (1980) administered the MRT at a comfortable listening level to naive listeners in order

to measure phoneme intelligibility. They got an overall error rate of 6.9% for synthetic speech as compared with a natural speech error rate of 0.6% using a different group of listeners. Performance for synthetic speech was somewhat better for initial consonants (4.6%) than for final consonants (9.3%). Nasals in the final position showed the highest error rate of 27.6%. The fricatives /θ/ and /ð/ also showed very high error rates.

In order to determine whether the performance for synthetic speech was inflated due to the forced-choice format of the MRT, the authors, as reported in Pisoni (1982), used an open free-response format. In this case subjects heard a single word but were required to write down the word they thought they heard, and to guess if necessary. The error rate for natural speech only increased from 0.6% to 2.8%. However, the error rate for synthetic speech increased dramatically from 6.9% to 24.6%, showing that the uncertainty of the open-response format caused a decrease in synthetic speech intelligibility. The authors also administered the same items from the MRT under different levels of masking noise and found that the perception of synthetic speech was more degraded under noise conditions than was the perception of natural speech.

In order to evaluate the recognition of words in sentences, Pisoni and Hunnicutt (1980) used two sets of test materials. The first set was composed of 100 Harvard Psychoacoustic Sentences which were meaningful and were hoped to replicate more normal listening conditions. The second set consisted of 100 meaningless sentences from the SNST developed at Haskins Laboratories by Nye and Gaitenby (1974).

These meaningless sentences supposedly allowed a much finer assessment of the contributions of acoustic-phonetic information to word recognition. These two different sentence tests were administered to two different groups of listeners. Correct word recognition on the Harvard sentences was quite good for synthetic speech with an overall mean of 93.2% correct versus 99.2% correct for natural speech with another group of listeners. However, as expected, due to the meaningless nature of the sentences, performance on the Nye and Gaitenby meaningless sentences was substantially worse with an overall mean of 78.7% correct for synthetic speech versus 97.7% correct for natural speech with another group of listeners. Pisoni and Hunnicutt (1980) noticed small but consistent learning effects between the first and second half of the test in both of the synthetic speech sentence tests.

Very similar results are noted when comparing the error rates as well as the percentage of words correct obtained in the studies of Pisoni and Hunnicutt (1980) regarding the MRT and the meaningless sentences with those of Nye and Gaitenby (1973, 1974) who used the same tests but a different synthesizer (see Tables III and IV). Mattingly (1980) states that even though the synthetic speech produced by these two speech synthesis systems is intelligible, synthetic speech places a much greater load on short-term memory than natural speech does in sentence recall tests.

Since most of the research in synthetic speech has been concerned with the intelligibility of words in isolation or words embedded in sentences, little is known about the comprehension and retention of

TABLE III

OVERALL AVERAGE ERROR RATES FOR SYNTHETIC VERSUS NATURAL SPEECH IN TWO STUDIES OF SYNTHETIC SPEECH INTELLIGIBILITY
(Taken from data as reported in Nye and Gaitenby, 1974 and Pisoni and Hunnicutt, 1980)

	MRT (N)	MRT (S)	SENTENCES (N)	SENTENCES (S)
Nye and Gaitenby (1974)	2.7%	7.6%	5.0%	22.0%
Pisoni and Hunnicutt (1980)	0.6%	6.9%	2.3%	21.3%

N = Natural speech
S = Synthetic speech

TABLE IV

PERCENTAGE OF WORDS CORRECT FOR SYNTHETIC VERSUS NATURAL SPEECH IN TWO STUDIES OF SYNTHETIC SPEECH INTELLIGIBILITY
(Taken from data as reported in Nye and Gaitenby, 1974 and Pisoni and Hunnicutt, 1980)

	MRT (N)	MRT (S)	SENTENCES (N)	SENTENCES (S)
Nye and Gaitenby (1974)	97.3%	92.4%	95.0%	78.0%
Pisoni and Hunnicutt (1980)	99.4%	93.1%	97.7%	78.7%

N = Natural Speech
S = Synthetic Speech

continuous passages of synthetic speech (Jenkins and Franklin, 1982). Pisoni and Hunnicutt (1980) evaluated the comprehension of continuous synthetic speech by having three groups of listeners answer multiple choice questions after listening to or reading passages from adult reading comprehension tests. One group listened to synthesized versions of the passage, while another group heard the natural version and a third group read the passages. The speaking rate was in excess of 180 words per minute, which closely approximates that of normal conversation or reading aloud (Pisoni, 1982). Pisoni and Hunnicutt (1980) found no significant differences between natural and synthetic speech. However, performance for the synthetic group improved by over 10% between the first and second half of the testing. Pisoni and Hunnicutt (1980) concluded that the listeners were able to adapt with little practice to relatively long passages of synthetic speech.

Jenkins and Franklin (1982) also investigated the effectiveness of continuous synthetic speech but they used speech synthesized by the OVE III speech synthesizer provided by Haskins Laboratories. The test material was very simple, consisting of a biography taken from the children's section of a newspaper. Intelligibility was tested by a dictation task involving three groups of listeners. The researchers concluded that synthetic speech was reasonably intelligible without practice (89%-95%) and with a little practice (90%-96%), it was virtually as intelligible as natural speech (96% to 98%). Although they found synthetic speech to be as intelligible and memorable as natural speech, the test materials consisted of very simple texts which may have hidden real differences which exist between natural and

synthetic speech.

Jenkins and Franklin (1982) stated that evaluations of synthetic speech are useful and timely in that they might uncover the possible reason for the well-known acoustic redundancy that is inherent in natural speech. They stated that synthetic speech usually lacks the acoustic redundancy (composed of some set of minimal phonemic cues) that is inherent in natural speech. Therefore, when comprehending continuous passages of synthetic speech, primary attention must possibly be given to identifying the individual phonemes which limits the amount of higher order processing (lexical, syntactic, and semantic) available to comprehend each sentence and relate it to the meaning of the whole passage.

Pisoni (1982) conducted a series of more sophisticated tests again using synthetic speech generated by the MITalk system. Using a lexical decision task in which listeners were to classify stimuli as a "word" or a "nonword", he found that performance was better for classifying natural speech items (98% correct) than it was for synthetic speech items (79% correct) and that the response times were faster for natural versus synthetic speech stimuli. In a subsequent experiment, listeners were asked to name or repeat "words" and "nonwords" and again it was found that the response accuracy as well as the response time was better for natural speech than it was for synthetic speech. Pisoni (1982) concluded that the present results of these two tests demonstrate that synthetic speech requires more cognitive processing time than natural speech to recognize or name words presented in isolation. He also concluded that the findings of

all studies so far regarding MITalk "demonstrate the existence of both perceptual and cognitive difficulties in perceiving synthetic speech signals."

Finally, Feustel, Luce, and Pisoni, as reported in Pisoni (1982), conducted a set of tests for short-term memory (STM) for word lists in synthetic and natural speech. They were trying to determine whether the differences in perceiving synthetic versus natural speech were due to "encoding and rehearsal processes in STM" or whether the differences were related to the difficulty of extracting important acoustic-phonetic information in the speech signal. The investigators used a memory preloading technique in which the subjects were asked to maintain zero, three, or six digits in STM while listening to word lists in synthetic and natural speech. The subjects then had to recall both the digits in correct serial order of presentation and as many of the words from the list as they could remember. The number of subjects who were able to correctly recall the digits decreased more rapidly for the synthetic lists versus the natural lists. These results indicate that the synthetic word lists interfered with the task of recalling digits more than the natural word lists did. In other words, that synthetic speech interferes in some way with the subject's ability to maintain information in STM.

The results of these studies regarding the identifying and naming of "words" and "nonwords" and the recall of words after memory preloading imply that the perception of synthetic speech may put increased processing demands on short-term memory (Jenkins and Franklin, 1982). Therefore, Pisoni (1982) states that we must be

careful when employing synthetic speech in conditions of high information loading such as aircraft cockpits, flight simulators, and computer-assisted instructional systems.

Some research has been done using digitized speech stimuli. Digitized speech is not true synthetic speech in that it is derived from a human source rather than generated by a computer. However, one study has been done using this highly intelligible form of speech which is related to the field of audiology and therefore, considered necessary for inclusion in this study. Digitized speech is highly intelligible and consequently may not place the stress on short-term memory that Pisoni (1982) talked about regarding formant synthesized speech. The drawback of this type of speech is that it requires huge amounts of storage, thereby necessitating a very limited vocabulary.

Cumming and McCorriston (1981) investigated the intelligibility of two types of digitized speech (Supertalker speech and Codec speech) for use in computer-assisted instruction (CAI) for beginning readers. The speech stimuli consisted of individual consonant sounds presented alone and in the initial and final position of short words. The subjects consisted of 47 children with a mean age of 6 years 9 months. The results demonstrated that Supertalker speech, with a mean identification performance of 78.2% correct as compared to 86.6% correct for natural speech is inadequate for use in CAI with beginning readers. However, Codec speech (mean 84.4% correct) gave performance closely approximating that of natural speech and therefore is considered acceptable for use with beginning readers.

A few studies have been concerned with using formant synthesized

speech in order to be able to manipulate various phonetically relevant acoustic aspects of the speech signal, such as formant frequency transitions and vowel and consonant length. Thus, it is possible to investigate the effects that these various spectral and temporal manipulations have on the intelligibility of speech and people with hearing loss.

Godfrey and Millay (1980) have begun to use synthetic speech to investigate the effects of hearing loss on the perception of particular types of speech sounds. They state that synthetic speech offers two advantages over natural speech. First, it allows for the precise control of the acoustic characteristics of the stimulus at every presentation. Second, acoustic cues, such as bursts and formant transitions in voiced stop consonants, can be isolated and separately tested, whereas in natural speech they may occur in combination. Godfrey and Millay (1980) used a computer-controlled formant synthesizer to synthesize several syllables (/ba-da/, /ba-wa/, /u-o/, /o-a/, /da-ya/) which differed in difficulty regarding the perception of the formant frequency transition cues. The perceptual tests were administered at 30 dB SL re SRT and at 100 dB SPL or just below the level of discomfort to normal hearing and sensorineural hearing-impaired subjects using a forced-choice identification task. The results confirmed the fact that sensorineural hearing impairment may cause difficulty in the identification of sounds cued by formant frequency transitions (Godfrey and Millay, 1980). The authors concluded that their synthetic speech stimuli "not only avoid the variability of naturally-spoken test materials, but also permit

investigation of the effects of hearing impairment on the perception of specific phonetically relevant acoustic cues."

With the same premise (ie., that synthetic speech allows for the controlled variation of perceptualy significant cues) Ginzel, et al. (1982a) conducted a related study. They synthesized 15 consonant-vowel (CV) stimuli which varied in second and third formant transitions in order to see if it would effect the categorical perception capabilities of older subjects with sensorineural hearing loss. The stimuli were presented via a single loudspeaker to three younger age groups with normal hearing and a group of older subjects with age-induced sensorineural hearing loss. The subjects were to identify the stimuli as belonging to one of three categories (/bae/, /dae/, /gae/). Response percentages were significantly lower and error frequency was high for the hearing loss group. Due to the fact that Ginzel, et al. (1982a) found poor speech sound identification in the group of older hearing subjects, they concluded that aging and age-induced sensorineural hearing loss are relevant factors in categorical speech perception. Since aging and age-induced senorineural hearing loss influence elderly people's ability to identify synthetic speech stimuli, it is possible that synthetic speech may prove useful in the audiological investigation of presbycusis.

Ginzel, et al. (1982b) used synthetic speech stimuli in order to investigate the effect of temporal factors on the process of categorical speech perception and found, as they had before with formant changes, that subjects with age-induced sensorineural hearing loss showed significantly lower response percentages than did younger

normal hearing subjects. They concluded that older people with signs of presbycusis perceive the temporal factors in auditory stimuli different than their younger normal hearing counterparts. Therefore, not only spectral cues, such as formant transitions, but also temporal factors, such as vowel and consonant length, are important in the process of categorical speech perception (Ginzel, et al., 1982a).

From the studies by Godfrey and Millay (1980) and Ginzel, et al. (1982a, 1982b) it is apparent that synthetic speech may provide a useful means to manipulate the speech signal in various ways which are not possible with natural speech. Synthetic speech can be manipulated in various ways in order to determine the effect that such factors as spectral and temporal cues have on various types of hearing impairment, especially that of presbycusis.

The intelligibility studies reported so far have all been concerned with the intelligibility of synthetic speech that has been generated using highly elaborate and very expensive formant resonance synthesizers. The primary goal of this research has been to identify various errors regarding the perception of synthetic speech in the hope of improving it for use other than the investigation of hearing impairment, such as in business and industry and reading machines for the blind. Therefore, the researchers did not use tests typically used for the audiological evaluation of hearing and did not carefully control or note presentation levels, as well as other variables in the testing situation. For instance, one variable that was not controlled in many of the studies cited concerned the subjects. Many studies used a different group of listeners for synthetic speech versus natural

speech. Although the potential uses and importance of synthetic speech has been well established along with the fact that synthetic speech has been shown to be fairly intelligible, only one research team has sought to determine the intelligibility of synthetic speech in speech audiometry. However, both of the studies done by this research team used Finnish speech.

Rahko, et al. (1979) investigated the intelligibility of synthetic Finnish speech generated by a portable, text-to-speech formant synthesizer, Synte 2, in order to determine whether synthetic speech could be understood and used in speech audiometry. They presented the Finnish speech words list at 30 dB SL with a word speed of 80 words per minute to 98 subjects. Some of the subjects had normal hearing while others had conductive or perceptive hearing defects. Rahko, et al. (1979) found that the mean discrimination score for normal hearing subjects was 100% for natural speech versus 70% for synthetic speech. The mean discrimination score for the conductive hearing loss group for synthetic speech was about the same (68.6%) as the normal hearing group, which is considered appropriate due to the fact that the discrimination test was administered at 30 dB SL. The mean discrimination score was slightly lower (62.2%) for the perceptive hearing loss group. See Table V for speech reception thresholds and discrimination scores for the various groups using natural and synthetic speech. In normal hearing subjects the discrimination score increased by 5% (from 70% to 75%) after 30 words of training. Decreasing the word speed caused an increase in speech discrimination. When the word speed was decreased to 50 words per minute, some subjects

TABLE V

THE DISCRIMINATION SCORES AND SPEECH RECEPTION THRESHOLDS IN NORMAL HEARING SUBJECTS AND SUBJECTS WITH DIFFERENT HEARING DEFECTS. WORD SPEED WAS 80 WORDS PER MINUTE. DISCRIMINATION SCORE WAS AT 30dB SL
(Rahko, et al, 1979)

	Normal hearing			Conductive hearing defect			Perceptive hearing defect		
	DC mean/SD (%)	SRT mean/SD (dB)	N	DC mean/SD (%)	SRT mean/SD (dB)	N	DC mean/SD (%)	SRT mean/SD (dB)	N
Synte 2	70.0/13.2	14.5/5.6	58	68.6/12.4	53.2/13.4	22	62.2/16.8	40.6/16.0	74
Synte 2 latter ear	75.3/10.0			73.4/11.5			68.3/15.4		
Finnish normal speech test	100	7.9/4.2	58	100	45.7/12.5	22	89.4/10.3	33.1/16.7	74

DC = Discrimination score; SRT = Speech reception threshold; N = Number of ears; SD = Standard deviation

DC = Discrimination score
SRT = Speech reception threshold
N = Number of ears
SD = Standard deviation

achieved 90% discrimination scores. Increasing the presentation level also caused an improvement in speech discrimination. Weaknesses in t, r, and j consonants were noted. On the basis of these discrimination test results, Rahko, et al. (1979) concluded that although Synte 2 has sufficient potential to synthesize speech, that the speech generated by this system could not be used for speech discrimination testing. It is possible that Rahko, et al. (1979) may have gotten such a low mean speech discrimination score (70%) for synthetic speech because of a rather low presentation level (30 dB SL).

In a second study, Rahko, et al. (1980) investigated the effect of various word speeds on the discrimination score of synthetic Finnish speech produced by Synte 2. 100 subjects with normal hearing were divided into four groups. Each group listened to synthetic speech words emitted by Synte 2 at 5-second intervals with rates corresponding to speech speeds of 50, 60, 70, or 80 words per minute. Each group listened to the same 150 words at 50 dB SL but at different speed of presentation. The mean discrimination score rose to over 80% after 100 words. The most dramatic increase in discrimination occurred between 75 and 100 words. After 100 words, a slight tendency towards a further increase in discrimination scores persisted. One group of subjects listened to 400 words and the best mean discrimination score of 84.4% occurred for the last 100 words. Therefore, discrimination scores did not increase greatly after subjects had listened to 75 to 100 words. The best mean discrimination score between 75 and 100 words was 91% at 60 words per minute. The results were about the same for all the word speeds tried. Therefore, word speeds ranging from 50-80 words per

minute didn't have much of an effect on the discrimination score. However, the number of correctly identified words reached its maximum at a presentation rate of 70 words per minute. Some weaknesses in the production of phonemes were noted. The phonemes /t, v, j, n, r/ were confused with other sounds and poorly perceived. The best consonants were /s, h, l, m/. The vowels were almost always heard correctly. Since subjects lacked the patience to continue listening to synthetic speech after 400 words, synthetic speech may be tiring or fatiguing to listen to after prolonged periods.

Rahko, et al. (1979) discussed the possibility of using synthetic speech for diagnosing central auditory problems. They found that increasing the word speed to 120 and 160 words per minute, which decreases the redundancy of the speech material, caused a decrease in the discrimination score. It is well known that tests which reduce the redundancy in natural speech material are useful in identifying central auditory impairments (Lundborg, et al., 1975; Snow, et al., 1977). Therefore, it may be possible to use synthetic speech tests in this manner. An added advantage of synthetic speech is that it can be speeded up without affecting the frequency area of the speech (Rahko, et al., 1979). Synthetic speech in its unmodified form lacks the acoustic redundancy that is inherent in natural speech (Jenkins and Franklin, 1982). Therefore, it might prove useful in identifying central auditory defects without various manipulations.

So far, these are the only studies (Rahko, et al., 1979; Rahko, et al., 1980) evaluating the intelligibility of synthetic speech for possible use in speech discrimination testing and they were done using

Finnish speech. According to Kiukaanniemi and Mattila (1980) there is a significant difference in the distribution of speech power between English and Finnish speech. This difference is thought to result from the difference in the vowel/consonant ratio between the two languages. The total frequency of occurrence of vowels in general English conversation is 38% versus 51% for Finnish speech. Therefore, there is a greater concentration of speech power in the low frequencies for Finnish speech than for English speech (Kiukaanniemi and Mattila, 1980). Kiukaanniemi and Mattila (1980) state that the differences in phonetic and linguistic structure of various languages may cause problems when trying to compare speech discrimination test results between different languages. Carhart (1953a) states "that the techniques of speech audiometry must be developed independently for every language in which they are used. . ." Therefore, it is also necessary to evaluate the intelligibility of synthetically produced speech discrimination word tests in English to see if they could be used as a possible replacement for present speech discrimination test materials.

Since 1978, there has been an increase in the number of commercially available voice output communication aids (VOCA'S) for the speech impaired or mute (Levinson and Kraat, 1984). One such device is the HC 120 Phonic Mirror Handivoice. Oggerino (1980) investigated the basic intelligibility of synthetic speech as produced by the HC 120. He tested 40 normal-hearing subjects and found that percent correct scores for eight spondee words ranged from 0% to 50% correct, while for eight PB monosyllable words scores ranged from 0% to 75% correct.

Phrase identification ranged from 72.5% to 97.5% correct. Oggerino (1984) concluded that the HC 120 was limited in its production of isolated words, but that it had good to excellent capabilities for the production of phrases and words organic to itself (i.e., in the brochure).

Williams, et al. (1980) compared the intelligibility of synthetic speech produced by the Versatile Portable Speech Prosthesis (VPSP) with the Handivoice, a new commercial unit which uses a Votrax synthesis chip. One group of ten subjects unfamiliar with listening to synthetic speech heard VPSP speech and another group heard the Handivoice speech. Synthetic speech stimuli consisted of sentences, a CNC word list, and one list from the Modified Rhyme Test (MRT). Intelligibility was defined as the percentage of words correct versus the total number of words presented. Results showed that the synthetic speech produced by the VPSP was superior in intelligibility to that produced by the Handivoice on all three intelligibility measures. Sentence intelligibility for the VPSP was 71% compared to 26% for the Handivoice. Word intelligibility for the MRT and CNC word list respectively was 68% and 44% for the VPSP versus 43% and 10% for the Handivoice. Levinson and Kraat (1984) stated that the results of this study by Williams, et al. (1980) indicated significant differences in speech intelligibility between these two systems and that both synthesizers were significantly reduced in intelligibility regarding natural speech, with the Handivoice being the most severly reduced.

Levinson and Kraat (1984) investigated the intelligibility of speech synthesized by the Votrax Personal Speech System (PSS) and the

Echo][speech synthesizer, two speech synthesizers used in communication aids for the speech impaired. Synthetic speech stimuli consisted of sixty-four sentences, each containing eight words, selected from the Assessment of Intelligibility of Dysarthric Speech (Yorkston and Beukelman, 1981). These sentences were administered to one adult listener unfamiliar with synthetic speech through an Apple][+ computer in a free-field condition in a sound treated room. Sentences were presented in two conditions for both synthesizers: a full sentence at a normal speaking rate, and a full sentence with 2 1/2 second pauses between each word in the sentence. Intelligibility scores were calculated as the number of words correctly identified for each sentence. The intelligibility scores differed significantly between the synthesizers. The synthetic speech produced by the Echo][was found to be 57% correct versus 72% correct for the synthetic speech produced by the Vortrax Personal Speech System in the no pause condition. Pause had a positive influence on the intelligibility for both synthesizers, especially for the Echo][. With the addition of pauses, the intelligibility score for the Votrax system increased from 72% to 81%, while the intelligibility score for the ECHO][system increased dramatically from 57% to 86%. Although their results indicated that pauses between words increased the overall intelligibility, Levinson and Kraat (1984) stated that their study needed to be repeated with a larger number of subjects before conclusive statements about differences in intelligibility and the effect of pause regarding these two speech synthesizers could be made.

The majority of studies on the intelligibility of synthetic

speech have used an expensive formant resonance synthesizer that has the most advanced synthesis rules and output means which are available to generate their synthetic speech materials. The speech produced by these synthesizers has been reported to be fairly intelligible. However, these intelligibility results should be viewed as the ultimate in synthetic speech intelligibility. Due to the high cost of the synthesizer, not many audiology clinics could afford to generate speech test materials in this manner. What is needed is a simpler and less expensive way to generate standardized test materials. Most of the speech synthesizers used in commercially available products for the disabled are low cost synthesizers with fewer rules and less memory (Levinson and Kraat, 1984). As of yet, these less expensive speech synthesizers have not been adequately researched. Initial tests of intellgibility of the more popular low cost, portable speech synthesizers reveal significantly reduced intelligibility scores (Levinson and Kraat, 1984). Rahko, et al. (1979) investigated the intelligibility of synthetic Finnish speech generated by a portable, formant speech synthesizer and found that the Finnish speech produced was not intelligible enough to be used for testing speech discrimination in its present form.

Due to recent advances in integrated circuit technology, there are now several single-chip synthesizers available which offer the ability to generate synthetic speech with a reliable, low-cost, portable speech synthesizer. However, before synthetic speech can be used with confidence in the audiological evaluation of hearing impairment, high quality levels of synthetic speech must be assured

(Damper, 1982). The recent advances in technology regarding speech synthesis systems has increased the need for reliable and economical ways to measure the intelligibility of the synthetic speech they produce (Voiers, 1983). Therefore, it is necessary to evaluate the intelligibility of a presently available speech synthesizer in order to determine whether the synthetic speech produced is intelligible enough to be used in speech discrimination testing in the audiological evaluation of hearing impairment. Linear predictive coding techniques are known to produce good quality speech (Canning, 1983). Therefore, it seems reasonable to assume that it would be possible to synthesize fairly intelligible speech using a portable, linear predictive coding (LPC) speech synthesizer.

This study is timely in that the addition of any newly automated technique is likely to be met with problems (Nye, et al., 1973). High quality synthetic speech is needed, but it is necessary to start preliminary research in order to determine the present quality of synthetic speech. It is necessary to identify the strengths and weaknesses of synthetic speech as well as the possible problems regarding its use in Audiology, so that necessary improvements can be made. According to Damper (1982), electronic companies will provide the proper technology, but is is up to us "clinical engineers" to discover how best to use it.

The ECHO][Speech Synthesizer

Synthetic speech stimuli were generated using the ECHO][speech synthesis system, manufactured by Street Electronics Corporation and based on Texas Instruments' TMS 5220 speech processor. The Echo][

speech synthesizer contains two programs for generating synthetic speech. The first program is called TEXTALKER. To begin synthesizing speech, the user runs the text-to-speech program called TEXTALKER. In order to generate speech using TEXTALKER, the word to be produced is simply typed into the computer and the Echo][speech synthesizer produces the word that was typed. TEXTALKER uses over 400 language and pronunciation rules to analyze a word. After analyzing the word, TEXTALKER converts the word into the appropriate phoneme and pitch codes and sends these codes to SPEAKEASY which is the second program for generating synthetic speech. It is a program which allows speech to be produced by using special phoneme codes. SPEAKEASY takes in the phoneme and pitch codes generated by TEXTALKER or typed in manually by the user and converts them to LPC parameters required to control the 5220 speech chip. The 5220 speech chip generates phonemes by passing digital codes for each phoneme through a digital filter which mathematically models the human vocal tract using linear predictive coding. Different phonemes may be produced by changing the LPC parameters for this filter. The phonemes are then converted from digital to analog signals, are amplified, and then sent out through the speaker or to the recording device (see Figure 5).

The resulting synthetic speech is remarkably intelligible according to the manufacturer. In their manual, the Street Electronics Corporation (1982) states that the text-to-speech system achieves approximately 90% accuracy in the pronunciation of English vocabulary. The ECHO][is a highly versatile speech synthesizer with the capability of providing an unlimited vocabulary. It contains 63

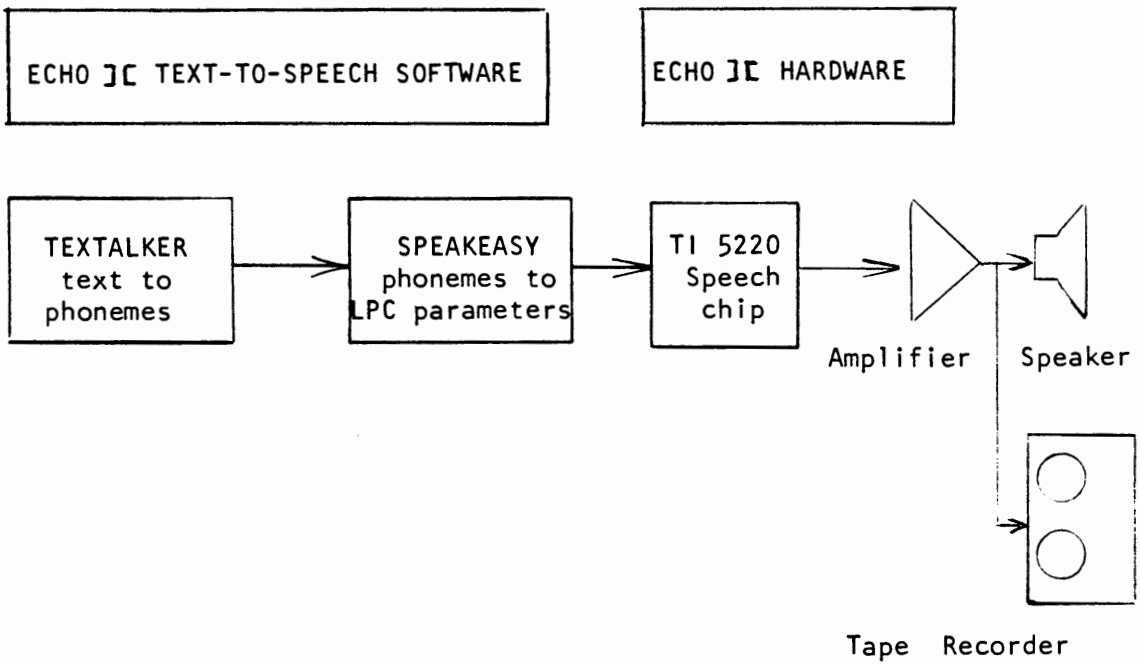

Figure 5. Schematic diagram of the ECHO][Speech Synthesis System.

different pitch levels and 15 different volume levels from which the user may choose. Words can be spoken monotonically (i.e., with even or unvaried pitch) or with intonation (i.e., variation in pitch) by using punctuation characters. For instance, a comma will create a pause, and a question mark will cause the pitch to rise, while a period will cause a lowering in pitch at the end of a sentence. The speech rate can be either fast or slow. Special features are also available to aid the blind or visually impaired user. And finally, the ECHO II speech synthesizer is a very simple and easy-to-use peripheral that is easily affordable. Therefore, there is no reason why it could not be easily added to the present equipment found in the audiological test suite.

The NU-6 Test

The traditional test for speech discrimination which is used in this country consists of word lists composed of 50 phonetically balanced monosyllable words (Carhart, 1965; Tillman and Olsen, 1973). The construction of the original word lists was done by the Harvard Psycho-Acoustic Laboratory (PAL) during World War II and was based on several criteria outlined by Egan in 1948 (Egan, 1948; Carhart, 1965; Hoode and Poole, 1980). These general criteria concerning the selection of words were that they must be familiar but not easy and that the word lists were to be phonetically balanced (Goetzinger, 1978).

However, the PAL PB-50 word lists were found to be inadequate due to the lack of familiarity of many of the test words and the poor standardization of the recordings. Also, they may not be phonetically balanced (Goetzinger, 1978). Therefore, Hirsh, et al. (1952) undertook

a modification of the PAL PB-50 word lists and constructed the well-known Central Institute for the Deaf (CID) W-22 test. They selected 120 words from a pool of 1000 PB-50 words and added 80 new words. Subsequently, the 200 words were arranged into four phonetically balanced lists, each containing 50 words and then recorded by Ira Hirsh (Goetzinger, 1978). However, there is considerable evidence that the CID W-22 word lists are too easy to be used in the differential diagnosis of various types of hearing loss (Burke, et al., 1965; Campbell, 1965; Goetzinger, 1978).

Due to the inadequacies in the phonetic balance of the PB-50 lists, Lehiste and Peterson (1959) developed a new monosyllabic word test. From a pool of 1263 consonant-nucleus-consonant (CNC) monosyllables, they prepared 10 lists of 50 words each. These lists were later revised (Peterson and Lehiste, 1962) and provided the foundation for the Northwestern University (NU) auditory test Nos. 4 and 6 (Davis and Silverman, 1970; Goetzinger, 1978).

The NU test No. 6 was developed to overcome some of the problems encountered with the previous monosyllable PB word tests. It is considered to be the most carefully prepared and thoroughly researched set of CNC word lists that has been published so far (Davis and Silverman, 1970). The NU-6 test consists of four lists of 50 words each which research has shown to have high interlist equivalence and test-retest reliability (Davis and Silverman, 1970, Rintelmann, et al., 1974; Hodgson, 1980).

Normative studies utilizing various recordings of the NU-6 test have found maximum discrimination scores to occur at approximately 32

dB SL (Rintelmann, et al., 1974; Beattie, et al., 1977). Rintelmann, et al. (1974) used his own recorded version of the NU-6 test and found that 96.8% to 98.0% of the words were correctly discriminated at 32 dB SL, depending on the word list, for 10 normal-hearing subjects. Beattie, et al. (1977) used the Auditec of St. Louis recording of the NU-6 test and found that 95% of the words were correctly identified at 32 dB SL for 24 subjects with normal hearing.

Although Rintelmann, et al. (1974) performed several experiments and found some small differences in the degree of difficulty between various NU-6 word lists, he concluded that when testing normal-hearing subjects all four lists are essentially equivalent for clinical purposes. Other research has also shown the NU-6 test to have high interlist equivalence (Davis and Silverman, 1970; Hodgson, 1980).

Summary

Synthetic speech is the production of human-like speech from a computer or speech synthesizer. There are three ways in which to synthesize speech; digitization, linear predictive coding, and formant synthesis. The last two methods actually produce synthetic speech, whereas digitization involves reproducing a speaker's voice.

The importance of synthetic speech has been proven by its many diverse applications. Synthetic speech is currently being used in various consumer products, in business and industry, and to provide aids for the handicapped. It is beginning to be implemented in the fields of education, speech pathology, audiology, and aural rehabilitation.

Several studies have been conducted in order to evaluate the

intelligibility of synthetic speech. Most of these studies used expensive formant resonance synthesizers to generate their synthetic speech materials and were concerned with the intelligibility of synthetic speech regarding its use in business and industry and in reading machines for the blind. The synthetic speech produced by these expensive synthesizers has been found to be fairly intelligible. The intelligibility of words in sentences has been reported to range from 77% to 85%. The intelligibility of phonemes has ranged from 89% to 91% for phonemes in words contained in sentences, while phoneme intelligibility was reported to be only 86% in individual words. The formant synthesizers were able to produce the vowel sounds with good accuracy. However, the nasals and especially the fricatives were rather poorly produced. Although less expensive, portable speech synthesizers are becoming available, preliminary tests show significantly reduced intelligibility results when compared with the expensive formant synthesizer intelligibility ratings.

There is a real need for a less expensive way to synthesize speech with a greater degree of intelligibility. Presently, speech synthesizers are becoming more and more affordable, and the intelligibility of synthetic speech is improving rapidly.

Synthetic speech offers many advantages to the audiological setting. Due to its greater degree of objectivity and lack of variability regarding natural speech, it could possibly be used to provide standardized speech test materials which could be comparable between clinics nationwide. Synthetic speech could also possibly prove useful in the evaluation of various types of hearing problems such as

presbycusis and central auditory hearing defects because its temporal and spectral aspects can be manipulated in ways that are not possible with natural speech stimuli.

PURPOSE

The dual purpose of this study was to develop two tape-recorded synthetic speech discrimination test tapes and assess the intelligibility of synthetic speech for possible use in audiological assessment. This study tried to determine whether or not synthetic speech was intelligible and if it would prove useful in speech discrimination testing.

Specifically, the experiment was directed toward the following objectives:

1) To evaluate the synthetic speech produced by the Echo][speech synthesizer using TEXTALKER and SPEAKEASY methods of generation and determine whether or not the Echo][speech synthesizer is useable for audiological assessment.
2) To compare the intelligibility of synthetic speech among normal-hearing listeners using TEXTALKER and SPEAKEASY methods of generation for monosyllable words in isolation (the NU-6 test).
3) To determine if there is a short-term learning effect.
4) To determine if there is a scrambling effect.
5) To determine the test-retest reliability.
6) To establish normative data with which further research might be compared.

CHAPTER III

METHODS

Subjects

Forty normal hearing volunteer subjects between the ages of 20 and 50 years were selected from students attending speech pathology and audiology classes at Portland State University. All subjects who took part in this study had normal hearing as measured under headphones using standard pure tone air conduction and speech discrimination testing procedures. Each subject had pure tone air conduction sensitivity better or equal to 15 dB HL (re ANSI, 1969) at the octave test frequencies between 250 and 4000 Hz in the test ear. Speech reception thresholds were within 5 dB HL of the average pure tone threshold for each ear. Also, speech discrimination scores in quiet for Campbell word lists (1965) were 96% or better in both ears for each subject when administered monaurally at 50 dB HL. The subjects who participated in this study reported no history of significant middle ear disease, familial history of hearing loss, or any other physiological or neurological impairment. All subjects reported no familiarity regarding the task of listening to synthetic speech.

Procedure

After a short case history questionnaire, subjects were seated in the audiological test suite and given an audiological assessment under

headphones using standard clinical procedures prior to inclusion in this study.

Only one ear per subject was used in this study due to the familiarity constraints imposed by the testing situation. The selection of which ear to test per subject was determined randomly by assigning numbers from a table of random numbers (Winer, 1962).

After the initial audiological evaluation, the subjects were read the following instructions through the headphones:

> You will be hearing two 50 word lists of monosyllable words in your ____ ear similar to the pretest stimuli you have just heard. The word lists are synthetic speech and may sound a little unnatural to you at first. Before hearing the two word lists you will be given five practice words in synthetic speech which are written on your response sheet. Your task is to listen to the words as presented in both lists and to say the word aloud and write down the word as you hear it on the response sheet provided. If you are unsure of a word, write down your best guess. Write down as much of the word as you can understand. If you cannot make out anything, draw a line through the numbered blank. There will be a ten second interval between words for you to record your response. You will also have a short rest period between the two synthetic word lists. Are there any questions before we get started?

Written responses were obtained in order to allow for the analysis of phonemic errors as well as to provide a more accurate account of the words as actually heard by the listener. A written response mode also precluded the possibility of the tester's hearing becoming an influential variable in the testing situation (Jerger, et al., 1968; Berger, 1971).

The test materials, consisting of four scramblings of the second NU-6 monosyllabic word list were generated in synthetic speech using two methods of generating synthetic speech called TEXTALKER and SPEAKEASY. The NU-6 word lists were chosen because they are composed of consonant-

nucleus-consonant words which would allow for more precise scoring and phonemic analysis. Also, the NU-6 word lists are considered to be the most carefully prepared and thoroughly researched CNC word lists published to date (Davis and Silverman, 1970). In addition, a full 50-word list was desired in order to make a comparison between the first versus the second half of the word list in the same ear in order to answer the question if there is a possible learning curve with synthetic speech as well as to maintain test effectiveness (Carhart, 1965; Nye and Gaitenby, 1973; Ainsworth and Miller, 1976; Rahko, et al., 1979). While all four lists are considered to be relatively equivalent, List 2 was chosen because of the findings of Rintelmann, et al. (1974). Rintelmann, et al. (1974) found that in general, list 1 was the most difficult and list 4 was the least difficult, while lists 2 and 3 were intermediate and essentially equivalent. They further stated that lists 2, 3, and 4 were essentially equivalent.

The order in which the four scramblings of the second NU-6 word list as well as the TEXTALKER and SPEAKEASY mode of presentation appeared was randomly determined. This was accomplished by assigning numbers from a table of random numbers (Winer, 1962).

Subjects were given a short rest period between the two synthetic speech discrimination tests in order to prevent the effects of fatigue from influencing the test results (Egan, 1948). In order to help overcome listening to the unnatural quality of synthetic speech, a short practice period consisting of five words from another synthetic word list generated by TEXTALKER preceded the speech discrimination tests. The practice period was short and did not provide the subjects

with any training regarding listening to synthetic speech. Also, the practice period did not provide training regarding listening to TEXTALKER speech versus SPEAKEASY speech because no feedback was given. This practice or adaptation period also allowed the subjects to become familiar with the presentation procedure and the written response mode.

The synthetic speech discrimination tests were administered at 60 dB HL. This level was chosen because it was consistent with that used by several researchers when testing with synthetic speech materials (Nye and Gaitenby, 1973; Nye and Gaitenby, 1974).

Instrumentation

The experimental test materials consisted of eight pre-recorded synthetic speech stimulus cassette tapes. Four tapes represented scramblings A through D of the second NU-6 word list as encoded using the TEKTALKER program, and the other four tapes represented scramblings A through D of the second NU-6 word list as encoded using the researcher's assignment of phoneme codes and the SPEAKEASY program. The test tapes were entitled TT2A, TT2B, TT2C, TT2D, and SE2A, SE2B, SE2C, SE2D, respectively.

These synthetic speech test tapes were developed using a 48K Apple][C personal computer with an ECHO][speech synthesizer board installed. The ECHO][speech synthesizer board is controlled using two different software programs called TEXTALKER and SPEAKEASY. The program called TEXTALKER accepts typed English words as input and converts these typed words into phoneme codes using a factory programmed set of rules containing 48 phonemes and diphthongs for

creating synthetic speech stimuli. The resulting phoneme codes are then sent to the second software program called SPEAKEASY. SPEAKEASY accepts phoneme codes directly from TEXTALKER or phoneme codes manually typed in by the user as input and converts these codes into machine level instructions on the Apple][C computer which in turn drive the speech synthesizer to voice the phoneme sounds through the output speaker of the synthesizer.

Each word of the second NU-6 word list was phonetically encoded using the two different techniques in order to test the intelligibility of the TEXTALKER and SPEAKEASY programs indpendently. A set of SPEAKEASY phoneme codes as well as pitch and duration codes are used to allow the ECHO][speech synthesizer to produce a word (for a complete description of the phoneme and pitch and duration codes used in this study, see Appendix A). In using the TEXTALKER program, the NU-6 word to be encoded was typed into the Apple][C computer using traditional orthography, and the resulting phoneme code that was assigned was written down. In using the SPEAKEASY program, the phoneme codes recommended by TEXTALKER for each word were used as a point of departure. Each word was listened to and these TEXTALKER codes were modified using the rules specified in the ECHO][manual until the researcher determined that any further modification of the phoneme codes would not improve the intelligibility of the synthesized word. Each word was synthesized in this manner, listened to and evaluated subjectively by the investigator, and then modified to improve intelligibility using additional phoneme, pitch, and duration codes if it was deemed necessary (see Appendix B for a complete list of the

stimulus test words and their appropriate encoded forms for both the TEXTALKER and SPEAKEASY programs). Also included in Appendix B are the international phonetic alphabet (IPA) symbols for each stimulus test word which were used in scoring subject's written responses (Wise, 1957).

Once the proper TEXTALKER and SPEAKEASY phoneme codes were established for each word and written down, a specially written software program called BUILDER was used to create files of phoneme codes for each scrambling of the NU-6 word list for both TEXTALKER and SPEAKEASY (8 files of 50 words composed of phoneme codes). These files were verified by the investigator to be accurate by reading the newly created file and redisplaying the phoneme codes on the screen of the computer and visually verifying that it matched the coding written down on the input sheet.

A second program, called PLAYER, was written which took each word in the word list and presented it at ten second intervals to the ECHO][speech synthesizer which in turn converted each word to speech output. In addition to synthesizing the 8 sets of 50 test words and the five practice words, the carrier phrase "You will write . . ." was also synthesized and pre-recorded into the test materials before each test word. This was done in order to prepare the listener for the test word and thereby reduce the variability in discrimination scores due to inattention or distractability (Egan, 1948). A 15 second 1000 Hz tone was also recorded at the beginning of each stimulus tape in order to ensure calibration.

In recording the stimulus tapes, the audio output of the ECHO][

speech synthesizer was connected to a cassette stereo tape recorder (Technics, Model RS-263AUS) through the audiometer (Maico, Model MA 24B) in order to allow for tone and synthetic speech calibration (see Figure 6). The audio output of the ECHO][speech synthesizer was jacked into the right auxiliary input of the audiometer. The right air output of the audiometer was jacked into the right input of the Technics cassette tape recorder. The mixer of the audiometer was set such that the signal from the ECHO][speech synthesizer was fed to the right channel of the tape recorder. The program PLAYER was then started on the Apple][C computer, and the output from the ECHO][speech synthesizer was recorded on magnetic audio cassette tape. This sequence was repeated 7 more times until all 8 test tapes were produced.

These 8 stimulus cassette tapes were played through the tape and accessory circuits (Technics, Model RS-263AUS) of a dual channel clinical audiometer (Maico, Model MA 24B) during the experimental testing situation (see Figure 7). The right channel output of the cassette recorder was patched into the right tape input jack of the audiometer. The stimulus was directed to the proper test ear using the mixer on the audiometer. All of the experimental tests took place in a double-walled acoustic suite (International Acoustics Corporation, Model 1403). All the test materials were presented through a pair of standard audiometric headphones (Telephonics, Model TDH-39) mounted in foam rubber cushions (Acoustic Research, Model MX 41/AR).

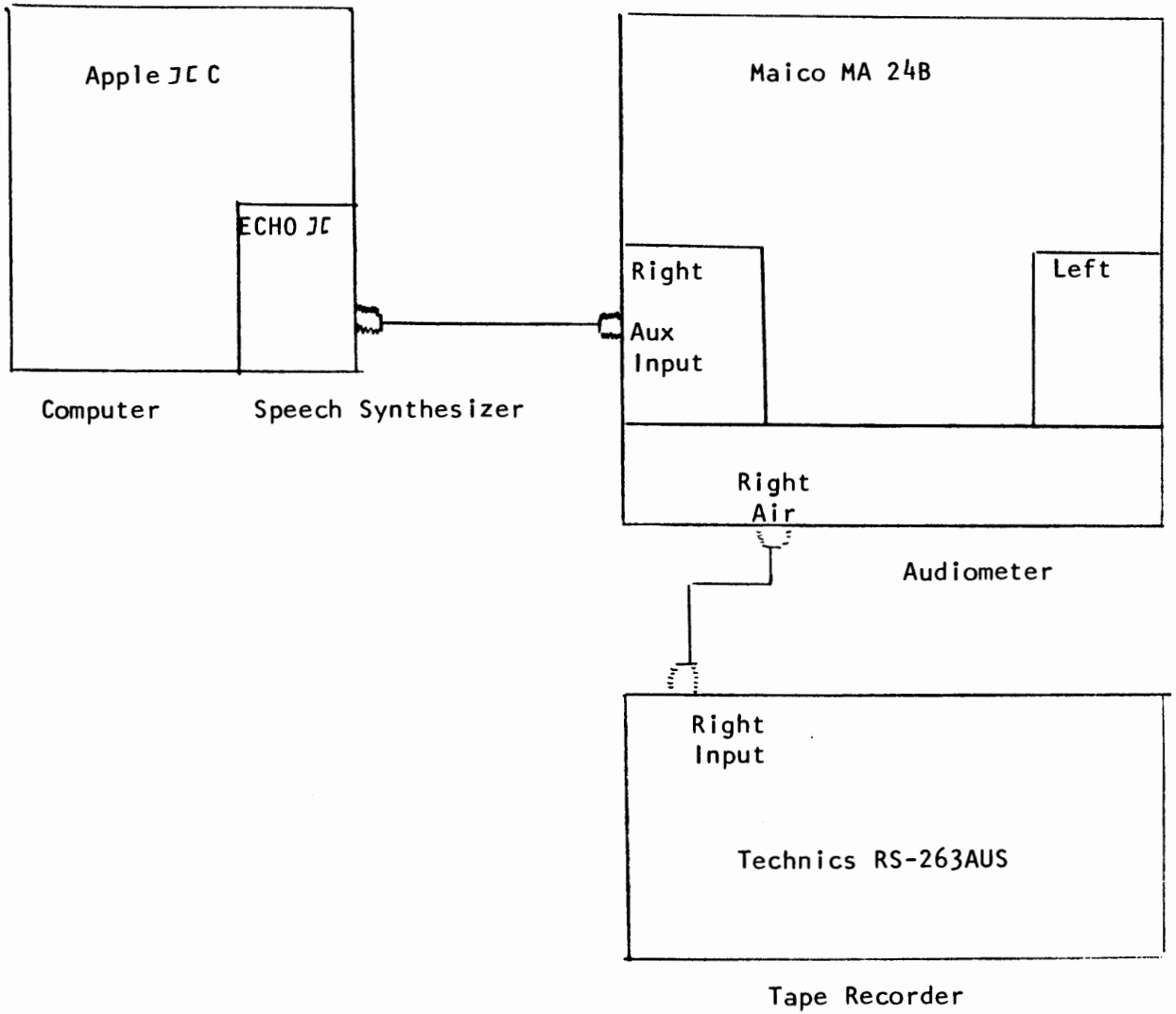

Figure 6. Schematic diagram for recording synthesized speech.

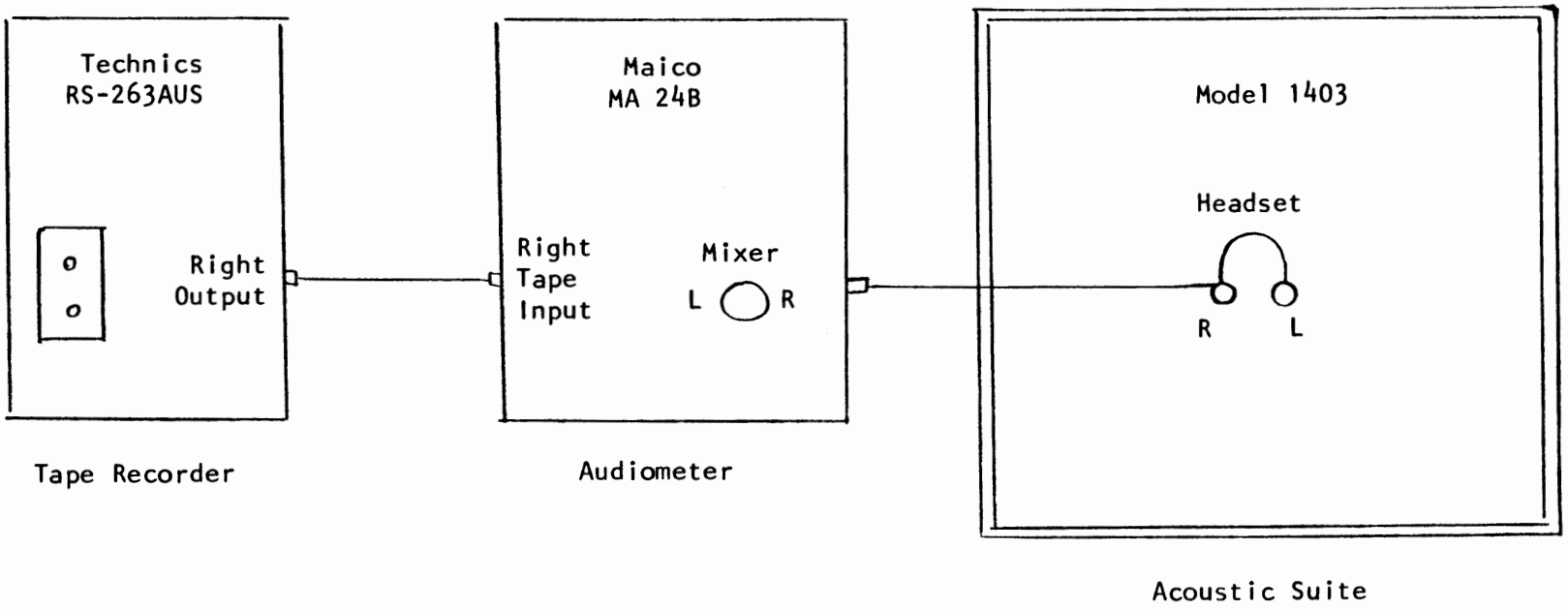

Figure 7. Schematic diagram of the testing instrumentation.

Calibration

The audiometer output at the earphones for both the left and right channels was electroacoustically calibrated before each experimental testing session to reflect current ANSI standards for pure tones (S3.6-1969) using a precision sound level meter (Bruel and Kjaer, Model 2203), a microphone (Bruel and Kjaer, Model 4132), and an artificial ear (Bruel and Kjaer, Model 4152). A prerecorded segment of a 1000 Hz pure tone was used to calibrate the speech circuit. The synthetic tape recorded materials along with the preceding calibration signals which were monitored at 0 dB HL on the VU meter were sent through the audiometer's tape and accessory circuits.

CHAPTER IV

RESULTS

The purpose of this study was to measure the intelligibility of synthetic speech generated by TEXTALKER and SPEAKEASY using normal hearing listeners and to determine whether or not the resulting synthetic speech generated by these two methods was usable for speech discrimination testing in audiological assessment. Data were collected and analyzed for forty normal-hearing adult subjects, 36 females and 4 males, who listened to two different scramblings of a 50 monosyllable word list, one scrambling generated by TEXTALKER and the other scrambling generated by SPEAKEASY. Subjects ranged in age from 20 to 50 years, with a mean age of 28.7 years.

Performance scores were obtained for each subject on both the TEXTALKER (TT) 50 monosyllable word list and the SPEAKEASY (SE) 50 monosyllable word list. Intelligibility was determined as the percentage of words correct out of the total number of words presented in each list for both TEXTALKER and SPEAKEASY. Table VI lists the mean, median, mode, standard deviation, and range of performance scores for TEXTALKER and SPEAKEASY individually and for both presentations combined. As can be seen in Table VI, the mean score for TEXTALKER (31.90%) was slightly below the SPEAKEASY mean score (33.55%) but this

TABLE VI

MEAN, MEDIAN, MODE, STANDARD DEVIATION, AND RANGE OF SCORES FOR TEXTALKER AND SPEAKEASY AND FOR BOTH PRESENTATIONS COMBINED FOR EXPERIMENTAL GROUP. N=40

PRESENTATION	MEAN	MEDIAN	MODE	S. D.	RANGE
TEXTALKER	31.90%	32%	34%	4.86	22%-40%
SPEAKEASY	33.55%	34%	38%	6.43	18%-46%
TT & SE COMBINED	32.73%	32%	34/38%	5.72	18%-46%

difference was not statistically significant (t=-1.30) at the 95% confidence interval (p>.05) (Winer, 1962). Figure 8 shows the generation effect between TEXTALKER and SPEAKEASY. Again, as can be seen from this figure, there is a very small increase in scores with the SPEAKEASY method of generating synthetic speech stimuli. A reasonably high Pearson Product Moment Correlation of +.69 demonstrates along with a small t value (-1.30) that the two tests are relatively similar and that a subject's performance on one test can be predicted from their performance on the other test.

Median and mode values were also higher for SPEAKEASY than TEXTALKER (Table VI). SPEAKEASY scores displayed a wider range by 10 points than TEXTALKER. For SPEAKEASY scores ranged from 18% to 46%. Table VII shows the actual percentage scores received and the total number of subjects who received each score for TEXTALKER and SPEAKEASY and for both presentations combined.

There was relatively little learning which took place when listening to synthetic speech generated by TEXTALKER and SPEAKEASY. Table VIII depicts the mean and standard deviation values for the first and second order presentations of SPEAKEASY and TEXTALKER. As can be seen from the values in Table VIII, mean scores increase slightly for the second order presentation for both TEXTALKER and SPEAKEASY but the increases are not statistically significant (t=-.74) at the 80% confidence interval (p>.20) (Winer, 1962). SPEAKEASY had a smaller increase in mean scores (.7%) than did TEXTALKER (1.2%). Figure 9 depicts the word discrimination scores obtained from the first versus the second order test for TEXTALKER and SPEAKEASY combined and shows

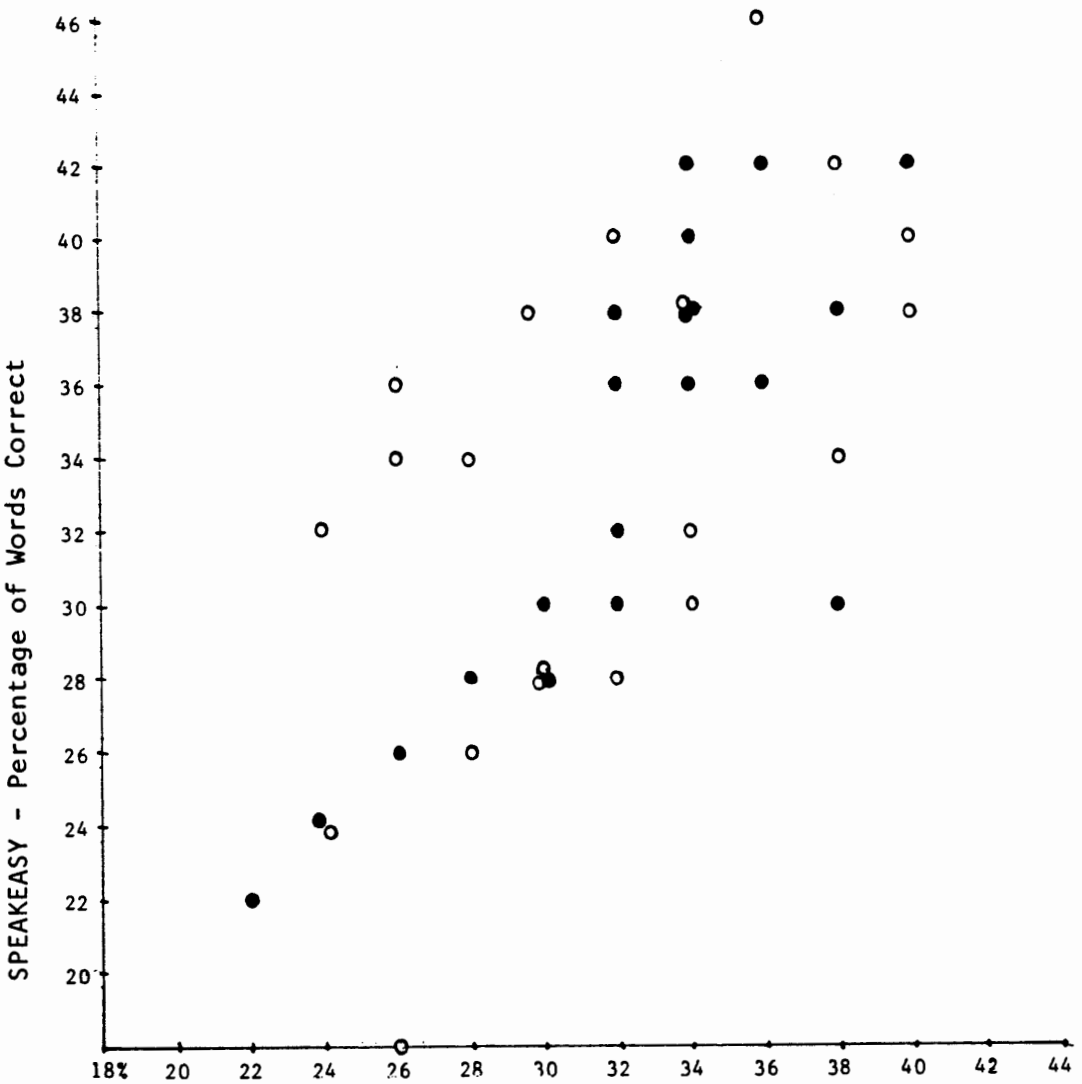

Figure 8. Word discrimination scores for TEXTALKER versus SPEAKEASY for each individual subject which shows the effect of the two different methods of generating synthetic speech for experimental group. N=40.

● = right ear o = left ear

TABLE VII

PERCENTAGE SCORES AND THE NUMBER OF SUBJECTS WHO RECEIVED EACH SCORE FOR TEXTALKER AND SPEAKEASY AND FOR BOTH PRESENTATIONS COMBINED FOR EXPERIMENTAL GROUP. N=40

PERCENTAGE SCORE	TEXTALKER	SPEAKEASY	TT & SE COMBINED
18%	0	1	1
20%	0	0	0
22%	1	1	2
24%	3	2	5
26%	4	2	6
28%	3	5	8
30%	5	4	9
32%	6	3	9
34%	8	3	11
36%	3	4	7
38%	4	7	11
40%	3	3	6
42%	0	4	4
44%	0	0	0
46%	0	1	1

TABLE VIII

MEANS AND STANDARD DEVIATIONS FOR FIRST AND SECOND ORDER PRESENTATION SCORES FOR TEXTALKER AND SPEAKEASY FOR EXPERIMENTAL GROUP. N=40

PRESENTATION	ORDER	MEAN	STANDARD DEVIATION
TEXTALKER	FIRST	31.3%	5.32
TEXTALKER	SECOND	32.5%	4.39
SPEAKEASY	FIRST	33.2%	6.03
SPEAKEASY	SECOND	33.9%	6.94

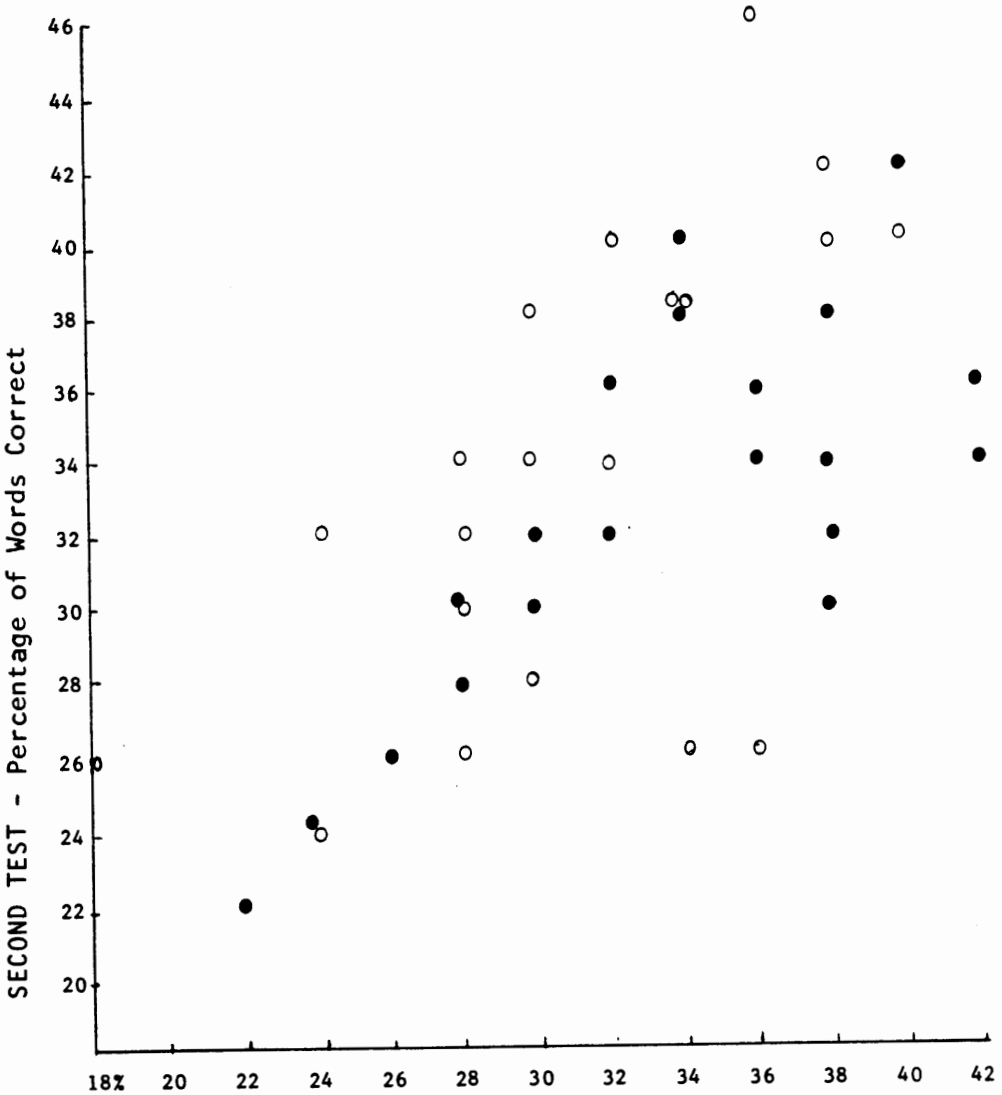

Figure 9. Word discrimination scores for the first test versus the second test for each individual subject which shows the order effect of the two tests for experimental group. N=40.

● = right ear o = left ear

that there was little evidence of learning. A relatively high Pearson Product Moment correlation of +.64 was obtained and again, coupled with a small t value (-.74) demonstrates that they are similar tests. Finally, another way to assess whether learning occurred was to evaluate each subject's response on the first versus the second half of the 50 monosyllable word test. Table IX lists the total number of words correctly identified for all subjects in the first versus the second half of the test for both TEXTALKER and SPEAKEASY. Again, there was no evidence of learning since an approximately equal number of words (roughly 50%) was correctly identified for the first versus the second half of the word test for both TEXTALKER and SPEAKEASY.

The various mean discrimination scores for each scrambling of the NU-6 word test were also evaluated for TEXTALKER and SPEAKEASY methods of presentation (see Figure 10). SPEAKEASY mean scores were slightly higher for all the scramblings with the exception of the A scrambling. The actual mean discrimination scores for each scrambling for both TEXTALKER and SPEAKEASY are reported in Table X.

In order to determine the reliability of the synthetic speech discrimination test generated by TEXTALKER and SPEAKEASY, t tests were performed and Pearson Product Moment Correlations were determined for the speech discrimination scores obtained for the five subjects. These values as well as mean percent correct scores and standard deviations are reported in Table XI. Due to the small subject sample size the Pearson R is underestimated.

Since word intelligibility was greatly reduced a phonemic analysis was not performed. Instead whole words were ranked according

TABLE IX

TOTAL NUMBER OF WORDS CORRECTLY IDENTIFIED FOR THE FIRST VERSUS THE SECOND HALF OF THE 50 MONOSYLLABLE WORD TEST FOR TEXTALKER AND SPEAKEASY FOR EXPERIMENTAL GROUP. N=40

PRESENTATION	WORDS 1-25	WORDS 26-50	ALL 50 WORDS
TEXTALKER	331(52%)	307(48%)	638
SPEAKEASY	326(49%)	345(51%)	671

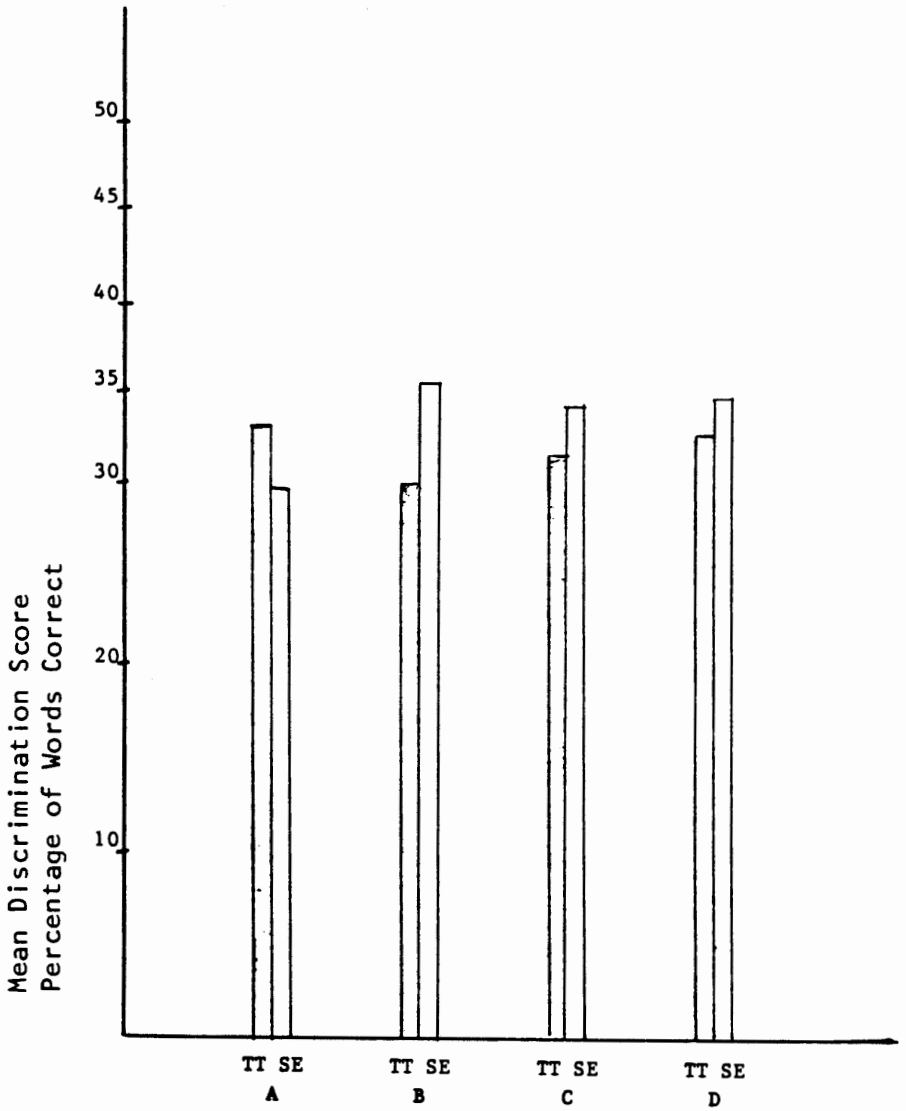

Figure 10. Mean discrimination scores for scramblings A through D of the 50 word test for TEXTALKER and SPEAKEASY for experimental Group. N=40.

TABLE X

MEAN DISCRIMINATION SCORES FOR SCRAMBLINGS A THROUGH D OF THE NU-6 WORD TEST FOR TEXTALKER AND SPEAKEASY FOR EXPERIMENTAL GROUP. N=40

PRESENTATION	A	B	C	D
TEXTALKER	33.0%	30.0%	31.8%	32.8%
SPEAKEASY	29.4%	35.8%	34.2%	34.8%

TABLE XI

MEANS, STANDARD DEVIATIONS, PEARSON PRODUCT MOMENT CORRELATIONS, AND T VALUES ON TEST/RETEST SPEECH DISCRIMINATION SCORES FOR TEXTALKER AND SPEAKEASY FOR EXPERIMENTAL GROUP. N=5

PRESENTATION	TEST/RETEST	MEAN	S.D.	t VALUE	PEARSON R
TEXTALKER	TEST	34.4%	4.33	-.79	.68
TEXTALKER	RETEST	36.8%	5.21	-.79	.68
SPEAKEASY	TEST	36.8%	5.21	-.67	.83
SPEAKEASY	RETEST	39.2%	5.93	-.67	.83
FIRST TEST	TEST	35.2%	5.21	-.76	.84
FIRST TEST	RETEST	38.0%	6.32	-.76	.84
SECOND TEST	TEST	36.0%	4.69	-.64	.71
SECOND TEST	RETEST	38.0%	5.09	-.64	.71

to those that were most intelligible to those that were least intelligible. Table XII lists the 50 monosyllable words in decreasing order of intelligibility from most to least intelligible for TEXTALKER, SPEAKEASY, and for TEXTALKER and SPEAKEASY combined. Several words such as "young", "room", "match", "rain", and "learn" were highly intelligible (89% - 100%) while other words such as "keg", "shack", "gaze", "goal", "said", "pad", and "shawl" were almost totally unintelligible (0% - 1%) for both TEXTALKER and SPEAKEASY combined. Word familiarity may have played a role in intelligibility and was not examined. The majority of words clustered around either high intelligibility or low intelligibility values with few words falling in between these two extremes. Word intelligibility fell off rather sharply around 75% and decreased to about 11% where a large portion of words grouped together with intelligibility slowly decreasing to 0%.

The number of whole words and individual phonemes correctly identified were tallied for each scrambling for both TEXTALKER (see Table XIII) and SPEAKEASY (see Table XIV). The total number of words correctly identified in SPEAKEASY for all scramblings (671) was slightly higher than those correctly identified in TEXTALKER (638) although the difference was not statistically significant. However, the total number of individual phonemes correctly identified for all three positions (I,M,F) was slightly higher for TEXTALKER (3789) than for SPEAKEASY (3698). The medial phonemes or vowels were the most intelligibile (79%, 77%) while initial (56%; 55%) and final consonants (54%; 53%) were equal in intellibilty for TEXTALKER and SPEAKEASY reapectively.

TABLE XII

PERCENTAGE CORRECT SCORES FOR NU-6 MONOSYLLABLE TEST WORDS
RANKED FROM MOST TO LEAST INTELLIGIBLE FOR TEXTALKER AND SPEAKEASY
AND FOR BOTH PRESENTATIONS COMBINED FOR EXPERIMENTAL GROUP

TEXTALKER AND SPEAKEASY COMBINED					TEXTALKER					SPEAKEASY				
RANK #	WORD	*C	*P	% CORRECT	RANK #	WORD	*C	*P	% CORRECT	RANK #	WORD	*C	*P	% CORRECT
1	YOUNG	80	80	100%	1	ROOM	40	40	100%	1	YOUNG	40	40	100%
2	ROOM	79	80	99%		YOUNG	40	40	100%	2	ROOM	39	40	98%
3	MATCH	77	80	96%	2	MATCH	39	40	98%	3	JUICE	38	40	95%
4	RAIN	75	80	94%	3	RAIN	38	40	95%		MATCH	38	40	95%
5	LEARN	71	80	89%		LEARN	38	40	95%	4	RAIN	37	40	93%
6	MILL	65	80	81%	4	WITCH	35	40	88%	5	LEARN	33	40	83%
	VOICE	65	80	81%	5	MILL	33	40	83%	6	TOOL	32	40	80%
7	WITCH	64	80	80%		VOICE	33	40	83%		MILL	32	40	80%
8	LOAF	60	80	75%	6	WHITE	32	40	80%		VOICE	32	40	80%
	CHAIR	60	80	75%	7	LOAF	29	40	73%	7	LOAF	31	40	78%
9	WHITE	55	80	69%		CHAIR	29	40	73%		CHAIR	31	40	78%
10	TOOL	54	80	68%	8	FAIL	25	40	63%	8	WITCH	29	40	73%
11	ROT	50	80	63%	9	CHIEF	24	40	60%	9	ROT	27	40	68%
12	CHIEF	48	80	60%		MERGE	24	40	60%	10	READ	25	40	63%
13	MERGE	46	80	58%	10	ROT	23	40	58%	11	CHIEF	24	40	60%
14	JUICE	39	80	49%	11	TOOL	22	40	55%	12	WHITE	23	40	58%
15	READ	34	80	43%	12	SOAP	18	40	45%	13	MERGE	22	40	55%
16	SOAP	32	80	40%	13	HATE	14	40	35%	14	NICE	19	40	48%
17	FAIL	27	80	34%	14	BOOK	11	40	28%	15	SOUTH	15	40	38%
18	NICE	24	80	30%	15	GIN	10	40	25%	16	SOAP	14	40	35%
19	SOUTH	23	80	29%	16	READ	9	40	23%	17	GIN	12	40	30%
20	GIN	22	80	28%		FAR	9	40	23%	18	FAR	8	40	20%
21	FAR	17	80	21%	17	SOUTH	8	40	20%	19	TON	6	40	15%
22	BOOK	16	80	20%	18	LIVE	7	40	18%		LORE	6	40	15%
23	HATE	15	80	19%	19	NICE	5	40	13%	20	DAB	5	40	13%
24	TON	11	80	14%		TON	5	40	13%		BOOK	5	40	13%
	LORE	11	80	14%		TURN	5	40	13%		THOUGHT	5	40	13%
25	LIVE	10	80	13%		LORE	5	40	13%		HAZE	5	40	13%
26	HAZE	9	80	11%	20	HAZE	4	40	10%	21	KEEP	4	40	10%
27	THOUGHT	7	80	9%	21	PICK	3	40	8%		NUMB	4	40	10%
28	PICK	6	80	8%		CALM	3	40	8%	22	PICK	3	40	8%
	KEEP	6	80	8%	22	KEEP	2	40	5%		WAG	3	40	8%
	CALM	6	80	8%		DEAD	2	40	5%		CALM	3	40	8%
	TURN	6	80	8%		WAG	2	40	5%		PIKE	3	40	8%
29	DAB	5	80	6%		PIKE	2	40	5%		LIVE	3	40	8%
	NUMB	5	80	6%		THOUGHT	2	40	5%	23	FAIL	2	40	5%
	WAG	5	80	6%		BITE	2	40	5%		HUSH	2	40	5%
	PIKE	5	80	6%	23	NUMB	1	40	3%		BOUGHT	2	40	5%
30	BITE	4	80	5%		JUICE	1	40	3%		BITE	2	40	5%
31	HUSH	3	80	4%		HUSH	1	40	3%	24	SAID	1	40	3%
	BOUGHT	3	80	4%		PAD	1	40	3%		HATE	1	40	3%
32	DEAD	2	80	3%		BOUGHT	1	40	3%		TURN	1	40	3%
	DEEP	2	80	3%		DEEP	1	40	3%		SHAWL	1	40	3%
33	SAID	1	80	1%	24	SAID	0	40	0%		DEEP	1	40	3%
	PAD	1	80	1%		DAB	0	40	0%	25	DEAD	0	40	0%
	SHAWL	1	80	1%		KEG	0	40	0%		KEG	0	40	0%
34	KEG	0	80	0%		SHACK	0	40	0%		SHACK	0	40	0%
	SHACK	0	80	0%		GAZE	0	40	0%		GAZE	0	40	0%
	GAZE	0	80	0%		SHAWL	0	40	0%		PAD	0	40	0%
	GOAL	0	80	0%		GOAL	0	40	0%		GOAL	0	40	0%

* C = # OF TIMES WORD WAS CORRECTLY IDENTIFIED.
* P = # OF TIMES WORD WAS PRESENTED.

TABLE XIII

NUMBER OF WHOLE WORDS AND INDIVIDUAL PHONEMES CORRECTLY IDENTIFIED FOR SCRAMBLINGS A THROUGH D OF THE NU-6 WORD TEST FOR TEXTALKER FOR EXPERIMENTAL GROUP. N=40.

SCRAMBLING	WORD	INITIAL	MEDIAL	FINAL	TOTAL_IMF/WD
A	165	292	381	273	946
B	150	287	399	263	949
C	159	278	395	266	939
D	164	267	410	278	955
TOTAL A-D	*638	*1124	*1585	*1080	3789

*-All total scores for whole words and initial, medial, and final phonemes are reported out of a possible correct total of 2000 points.

TABLE XIV

NUMBER OF WHOLE WORDS AND INDIVIDUAL PHONEMES CORRECTLY IDENTIFIED FOR SCRAMBLINGS A THROUGH D OF THE NU-6 WORD TEST FOR SPEAKEASY FOR EXPERIMENTAL GROUP. N=40

SCRAMBLINGS	WORD	INITIAL	MEDIAL	FINAL	TOTAL_IMF/WD
A	147	276	381	261	918
B	179	265	386	271	922
C	171	287	383	270	940
D	174	268	385	265	918
TOTAL A-D	*671	*1096	*1535	*1067	3698

*-All total scores for whole words and initial, medial, and final phonemes are reported out of a possible correct total of 2000 points.

CHAPTER V

DISCUSSION

This study employed synthetic speech generated by the ECHO][speech synthesizer using the two programs, TEXTALKER and SPEAKEASY, in order to measure the present intelligibility of synthetic speech among normal-hearing listeners for possible use in speech discrimination testing. The results indicated that synthetic speech as generated by the ECHO][speech synthesizer is significantly reduced in intelligibility and at this time may not be used as a substitute for normal human speech when testing speech discrimination ability of normal hearing listeners in audiological assessment.

The mean performance scores for TEXTALKER (31.90%) and SPEAKEASY (33.55%) are considerably below those obtained by previous researchers (75.4% to 93.2%) using expensive parallel formant synthesizers (Nye and Gaitenby, 1973; Nye and Gaitenby, 1974; Ingemann, 1978; Pisoni and Hunnicutt, 1980; Pisoni, 1982). However, most of these studies were either concerned with the intelligibility of words in sentences which would allow for the increase in scores because of the higher intelligibility afforded in sentences due to contextual cues, or were concerned with the intelligibility of words using the Modified Rhyme Test (MRT) which would also inflate scores due to the forced choice format. The only intelligibility test of synthetic speech which did

not fall into these two categories received the lowest of all intelligibility scores reported in the literature regarding expensive parallel formant synthesizers. Pisoni (1982) found word intelligibility scores of 75.4% using an open free-response word test. Rahko, et al. (1979) generated Finnish speech using a portable, formant speech synthesizer and reported a mean discrimination score of 70% which is also well above the mean scores reported in this study for both TEXTALKER and SPEAKEASY. However, this study was done using Finnish speech which has more vowels than English speech which may have contributed to greater intelligibility scores.

Although the mean scores for TEXTALKER and SPEAKEASY show the intelligibility of synthetic speech produced by the ECHO][speech synthesizer to be significantly reduced in comparison to the intelligibility of synthetic speech utilizing expensive formant resonance synthesizers, the mean scores reported in this study compare favorably to those of other researchers using the more affordable and commercially available speech synthesizers. Williams, et al. (1980) found CNC word intelligibility to be 44% for the Versatile Portable Speech Prosthesis (VPSP) and only 10% for the Handivoice. Levinson and Kraat (1984) found word intelligiblity in sentences to be 57% for the ECHO][speech synthesizer in the no pause condition. This score is a little above that obtained in the present study for TEXTALKER and SPEAKEASY using the ECHO][speech synthesizer, but again intelligibility scores were derived from words contained in sentences rather than from isolated monosyllable words.

The mean performance score for SPEAKEASY (33.55%) was slightly

higher than the TEXTALKER mean performance score (31.90%) but the difference was not statistically significant. This difference was probably due to the fact that the words were generated differently for each method.

In TEXTALKER, the words were typed into the computer using conventional orthography. Exact spelling for each word was used with no phonemic or sound spellings substituted for exception words. This procedure caused a severe limitation for the production of the word "juice" which was distorted and pronounced /dʒɔɪs/. The poor identification rate of this word was evidenced by the fact that it was correctly perceived 38/40 presentations for the SPEAKEASY method of generation and only 1/40 times for the TEXTALKER method of generation. The exact spelling procedure also caused the words "live" and "read" to be pronounced /laɪv/ and /rɛd/ for TEXTALKER instead of /lɪv/ and /rid/. However, they were accepted as correct when reported by subjects for the TEXTALKER presentation and consequently did not make a difference in scores between the two methods.

In SPEAKEASY, there was much more freedom in generating a word. If the TEXTALKER pronunciation seemed deficient, the researcher sometimes spelled words phonemically or sound spelled them. If the word produced using this procedure did not seem adequate, other techniques, such as variations in pitch and duration were used as well as inserting additional phonemes. One technique which backfired, actually causing more words to be missed in SPEAKEASY, was the addition of the schwa vowel /ə/ after some final stop plosive consonants. This was done in the hope of elongating the final consonant sound, thereby

making identification easier, but in actuality it caused more confusion and the word was ultimately missed. In some cases it caused the word to be perceived as being composed of two syllables. For example, the words /dɛd/ and /dæb/ were perceived as /dætə/ by some subjects.

In this study, there was little or no evidence of a short-term learning effect. Subjects' performances did not improve with presentation time for the first versus the second 50 monosyllable word test or from the first versus the second half of each individual 50 monosyllable word test for both TEXTALKER and SPEAKEASY. What was noticed on retrospect when grading subjects' answer sheets was that there were areas on each 50 word test where certain words were identified correctly. These areas varied with each scrambling which showed that certain words were usually correctly identified and other words were usually missed irregardless of presentation time and that the words did not become more intelligible with adaptation.

There was no effect noticed between the different scramblings of the NU-6 word test. The relationship between the mean scores for TEXTALKER and SPEAKEASY was stable with the exception of the A scrambling for which no reason was noted.

The test-retest reliability appeared to remain relatively stable for TEXTALKER and SPEAKEASY. There was a slight increase in mean scores for the second versus the first presentation for both TEXTALKER and SPEAKEASY but this slight increase appeared to be consistent with what would be expected when administering two different scramblings of the same 50 monosyllable word test on repeated occasions.

Although the total number of whole words correctly identified was

slightly higher for SPEAKEASY (671) than TEXTALKER (638), the total number of individual phonemes correctly identified was slightly higher for TEXTALKER (3789) than SPEAKEASY (3698). This may have been due to the researcher's manipulation of the rules in SPEAKEASY. In other words, the variations imposed by the researcher may have improved the intelligibility of whole words but actually decreased the intelligibility of individual phonemes for the SPEAKEASY method of presentation.

The medial phonemes, or vowels, were found to be the most intelligible (79%;77%) with initial (56%;55%) and final consonants (54%;53%) equal in intelligibility for TEXTALKER and SPEAKEASY respectively. The data from previous researchers regarding the intelligibility of initial and final phonemes is not in agreement with that found in the present study (Nye and Gaitenby, 1974; Pisoni and Hunnicut, 1980). For instance, Nye and Gaitenby (1974) found that word final consonants were more intelligible than initial consonants for words in the Modified Rhyme Test (MRT) and in a meaningless sentence test. Conversely, Pisoni and Hunnicutt (1980) administered the MRT and found that initial consonants (4.6% error rate) were slightly more intelligible than were final consonants (9.3% error rate). Although these two studies conflict regarding the intelligibility of initial and final phonemes, the data from this present study should not be expected to agree with either study due to the different synthesizers and word and sentence tests used.

The final aim of this study was to provide normative data with which further research regarding synthetic speech might be compared.

Although synthetic speech generated by the ECHO][speech synthesizer using TEXTALKER and SPEAKEASY methods of generation is significantly reduced in intelligibility compared to normal human speech, there is still a possibility that with future refinements, it may one day be used as a substitute for normal human speech in speech discrimination testing.

There may be several reasons for the reduced intelligibility results. First of all, an open free-response monosyllable word test is a very difficult test situation in that perception lies solely on accurate individual speech sound identification and is not limited by alternative word choices as found in closed-response tests or enhanced by contextual cues as found in sentence tests. Another reason for the reduced intelligibility scores may be due to the fact that synthetic speech is tiring to listen to for prolonged periods of time. Even with the presentation of only two 50 monosyllable word lists, some subjects reported fatigue, restlessness, and inattentiveness. A major reason for the poor intelligibility results was due to the ECHO][speech synthesizer itself. Not all speech sounds can be easily or adequately synthesized. The voiceless fricatives /f,θ,ʃ,h/ were very poorly produced and in actuality sounded like noise bursts. In addition, the rules available in the SPEAKEASY program for improving synthetic speech production were very limited and rather ineffective.

Conclusion

The purpose of this study was to develop two tape-recorded synthetic speech discrimination test tapes and assess their

intelligibility in order to determine whether or not synthetic speech was intelligible and if it would prove useful in speech discrimination testing. Four scramblings of the second NU-6 monosyllable word list were generated by the ECHO II speech synthesizer using two methods of generating synthetic speech called TEXTALKER and SPEAKEASY. These stimuli were presented in one ear to 40 normal-hearing adult subjects, 36 females and 4 males, at 60 dB HL under headphones. Each subject listened to two different scramblings of the 50 monosyllable word list, one scrambling generated by TEXTALKER and the other scrambling generated by SPEAKEASY. The order in which the TEXTALKER and SPEAKEASY mode of presentation occurred as well as which ear to test per subject was randomly determined.

The mean performance scores for TEXTALKER and SPEAKEASY demonstrated that the intelligibility of synthetic speech produced by the ECHO II speech synthesizer was significantly reduced in comparison to the intelligibility of normal human speech and to synthetic speech produced by expensive formant resonance synthesizers. However, the mean performance scores reported in this study compared favorably to those of other researchers who used the more affordable and commercially available speech synthesizers.

The mean performance score for SPEAKEASY was slightly higher than the mean performance score for TEXTALKER but the difference was not statistically significant. There was relatively little learning which occurred when subjects listened to synthetic speech generated by TEXTALKER and SPEAKEASY. Also, there was no effect noticed for the different scramblings of the second NU-6 word list for TEXTALKER and

SPEAKEASY. The test-retest reliability appeared to remain relatively stable for TEXTALKER and SPEAKEASY. Several words such as "young", "room", "match", "rain", and "learn" were highly intelligible while others such as "keg", "shack", "gaze", "goal", "said", "pad" and "shawl" were almost totally unintelligible for both TEXTALKER and SPEAKEASY. The medial phonemes or vowels were most intelligible while initial and final consonants were approximately equal in intelligibility. The voiceless fricatives /f,θ,ʃ,h/ were very poorly produced.

These results suggest that synthetic speech as generated by the ECHO II speech synthesizer using TEXTALKER and SPEAKEASY methods of generation is not of sufficient intelligibility at this time to be used as a substitute for normal human speech in speech discrimination testing. It is further concluded that synthetic speech is not intelligible for single syllable words in isolation and its role in Audiology is questionable at the present time. However, speech synthesizers are continually improving and new synthesizers are being produced. While this study indicates that synthetic speech is not of adequate intelligibility to be useful in speech discrimination testing at the present time, future improvements in speech synthesizers may make investigations necessary in order to re-evaluate the possibility of using synthetic speech in speech discrimination testing. This researcher feels that we are on the threshold of speech synthesis and that within a few years time, it will be of sufficient intelligibility to prove useful in testing speech discrimination and will subsequently be integrated into the audiological test setting.

Implications for Future Research

The results of this study suggest a number of possible areas for future research. First of all, due to the reduced intelligibility results found with ECHO II synthetic speech stimuli in open free-response monosyllable word tests, it would be useful to investigate the intelligibility of synthetic speech using spondaic words or sentences as test stimuli. They would certainly provide a less difficult test stimulus than monosyllable words and intelligibility results would be expected to be greatly increased. It is possible that synthetic speech materials in multisyllable word and sentence forms would be the first to be implemented in the audiological test setting. It might be of added benefit to devise a multi-dimensional scoring system so that words and other responses could be qualitatively evaluated rather than marked as right or wrong. In the present scoring system, words which differed by only one phoneme (i.e., pɑrp/pɑrk) and words which were totally acoustically unrelated (i.e., gæg/dɛd) were both counted as word wrong errors. In these cases, a phonemic analysis for type of errors may provide additional information and identify problem phonemes needing further refinement.

It would also be useful to look at other synthesizers which are presently available. In the last few years, there have been tremendous advances in the quality of speech synthesis systems and new ones are continually being introduced. There are other speech synthesizers which are commercially available at the present time that may offer increased intelligibility results over those obtained with the ECHO II speech synthesizer. One such speech synthesizer is the Votrax

Type-'N-Talk text-to-speech synthesizer which employs the SC-01 phoneme-based speech synthesizer chip. The Votrax speech synthesizer is based upon formant synthesis techniques instead of linear predictive coding techniques used in the ECHO][speech synthesizer, which may yield improved intelligibility results.

It might be suitable to wait for a sufficient period of time until new ways to program and synthesize speech become available, thereby increasing synthetic speech intelligibility. Ingemann (1978) is reported to be working on a new program for synthesizing speech which is based on internal rules which give greater weight to the syllable as a unit of speech production. Other alternative ways to program speech synthesizers will come eventually and with them will comes a subsequent increase in the intelligibility of synthetic speech.

Due to the increased versatility of synthetic speech over that of natural speech, it may be possible to use synthetic speech materials in several different audiological testing situations and with several different hearing populations. For instance, it is well known that tests which reduce the redundancy in natural speech materials are useful in the identification of central auditory disorders (Lundborg, et al., 1975; Snow, et al., 1977). Synthetic speech can be speeded up without affecting the frequency area of the speech which would cause severe transient, consonant, and vowel pitch difficulties with natural speech stimuli (Rahko, et al., 1979). Therefore, it may be possible to produce a sensitized speech test for the diagnosis of central auditory problems using synthetic speech in this manner, due to the decreased redundancy in the speeded up synthetic speech stimuli.

At present, synthetic speech represents a difficult listening situation as evidenced by the reduced intelligiblity results of this and other studies using commercially available, portable speech synthesizers (Williams, et al., 1980; Levinson and Kraat, 1984). Jenkins and Franklin (1982) hypothesize that synthetic speech in its unmodified form, lacks the acoustic redundancy (composed of some set of minimal phonemic cues) which is in natural speech. Therefore, when listening to synthetic speech, the listener must give primary attention to phonemic identification which in turn limits the amount of higher order processing (lexical, syntactic, and sematic) which is possible when listening to natural speech. Pisoni (1982) also concludes that synthetic speech requires more cognitive processing time than natural speech and therefore, demonstrates perceptual and cognitive difficulties involved with the perception of synthetic speech. Due to the decreased redundancy noted with synthetic speech in its present form, it may prove useful in identifying central auditory defects without the various manipulations noted above by Rahko, et al. (1979). Therefore, at a future time when synthetic speech intelligibilty is improved and normative data established, it might be a good idea to test a group of normal-hearing listeners and a group of listeners with known central auditory impairments, in order to determine if there is a diagnostically significant difference in discrimination scores between the two groups for synthetic versus natural speech stimuli. It might also be of interest to obtain normative data on other hearing populations such as children and geriatric listeners in order to determine the relative diagnostic ability of synthetic speech materials

regarding age related effects such as maturation (i.e., unfinished linguistic development) and presbycusis.

Finally, this researcher is of the opinion that speech synthesis systems will continue to be improved to the point where the intelligibility of synthetic speech will be sufficient to be useful as a substitute for normal human speech in the audiological test setting. Before this happens, much must be done to improve the quality and intelligibility of presently available commercial speech synthesizers. That is why intelligibility studies such as this one are useful and timely in order to identify the strengths and weaknesses of synthetic speech as well as the possible problems regarding its use in Audiology so that the necessary improvements can be made by the manufacturers. According to Damper (1982), electronic companies will provide the proper technology, but it is up to us "clinical engineers" to discover how best to use it. If the intelligibility of synthetic speech can be perfected, synthetic speech will be of great benefit to the audiological test situation. Repeatable results, which will be comparable between clinics nationwide, will be obtained as a result of synthetic speech stimuli that are identically generated and administered automatically.

REFERENCES

Ainsworth, W.A., "A Method of Estimating Speech Synthesizer Parameters by Temporal Analysis of Waveforms." Int. J. Man-Machine Studies, 3(4):339-49, 1971.

Ainsworth, W.A., Mechanisms of Speech Recognition. Oxford: Pergamon Press Ltd., 1976.

Ainsworth, W.A., and Miller, J.B., "Allophonic Variations of Stop Consonants in a Speech Synthesis-by-Rule Program." Int. J. Man-Machine Studies, 8(2):159-68, 1976.

Allen, J., "Linquistic-Based Algorithms Offer Practical Text-to-Speech Systems." Speech Technol., 1(1):12-6, 1981.

American National Standards Institute, Specifications for Audiometers. ANSI 53.6-1969, New York: American National Standards Institute, Inc., 1970.

Babu, B.N.S., and Preuss, R.D., "A Note on Complexity Reduction for Linear Predictive Speech Synthesis." IEEE Trans. Acoust., Speech and Signal Process., ASSP-30(3):516-9, 1982.

Beattie, R.C., Edgerton, B.J., and Svihovec, D.V., "A Comparison of the Auditec of St. Louis Cassette Recordings of NU-6 and CID W-22 on a Normal-Hearing Population." JSHD, 42:60-4, 1977.

Berger, K.W., "Speech Audiometry." IN: Audiological Assessment, Rose, D.E., (Ed.), Englewood Cliffs, New Jersey: Prentice-Hall, Inc., 1971.

Bess, F.H., "Clinical Assessment of Speech Recognition." IN: Principles of Speech Audiometry, Konkle, D.F., and Rintelmann, W.F., (Eds.), Baltimore: University Park Press, 1983.

Brandy, W.T., "Reliability of Voice Tests of Speech Discrimination." JSHR, 9:461-5, 1966.

Burke, K.S., Shutts, R.E., and King, W.P., "Range of Difficulty of Four Harvard Phonetically Balanced Word Lists." Laryngoscope, 75(2):289-96, 1965.

Campbell, R.A., "Discrimination Test Word Difficulty." JSHR, 8:13-22, 1965.

Canning, R.G., "Is 'Voice' in Your Future Systems?" EDP Analyzer, 1-16, August, 1983.

Carhart, R., "Basic Principles of Speech Audiometry." Acta Otolaryngol., 40:62-71, 1953a.

Carhart, R., "Speech Audiometry in Clinical Evaluation." Acta Otolaryngol., 41:18-42, 1953b.

Carhart, R., "Problems in the Measurement of Speech Discrimination." Arch Otolaryngol., 82:253-60, 1965.

Ciarcia, S., "Use ADPCM for Highly Intelligible Speech Synthesis." Byte, 35-49, June, 1983.

Cohen, J.R., "Segmenting Speech Using Dynamic Programming." JASA, 69(5):1430-8, 1981.

Cumming, G., and McCorriston, M., "Evaluation of Computer Speech for Use with CAI for Young Children." Journal of Computer-Based Instruction, 8(2):22-7, 1981.

Curry, E.T., and Cox, B.P., "The Relative Intelligibility of Spondees." JAR, 6:419-24, 1966.

Damper, R.I., "Speech Technology - Implications for Biomedical Engineering." J. Med. Eng. and Technol., 6(4):135-49, 1982.

Darrow, B., "Advances in Speech Technology Give Voice to Consumer Products, Motor Abilities to the Disabled." Design News, October, 1983, pp. 76-80.

Davis, A., and Berman, A., "Advanced Data Acquisition Aids the Handicapped." Computer Design, :121-31, 1983.

Davis, H., and Silverman, S.R., Hearing and Deafness. New York: Holt, Rinehart and Winston, 1970.

Edward, J.A., "Recent Work at JSRU on Speech Synthesis by Rule." Paper presented at the IEE Colloquium on 'The Computer Generation of Speech'." London, England, pp.1-4, 1981.

Egan, J.P., "Articulation Testing Methods." Laryngoscope, 58:955-91, 1948.

Elkins, E.F., "Analysis of the Phonetic Composition and Word Familiarity Attributes of CNC Intelligibility Word Lists." JSHD, 35(2):156-9, 1970.

Elliott, L.L., Longinotti, C., Meyer, D., Raz, I., and Zucker, K., "Developmental Differences in Identifying and Discriminating CV Syllables." JASA, 70(3):669-77, 1981.

Elovitz, H.S., Johnson, R., McHugh, A., and Shore, J.E., "Letter-to-Sound Rules for Automatic Translation of English Text to Phonetics." IEEE Trans. Acoust., Speech and Signal Process., ASSP-24(6):446-58, 1976.

Elpern, B.S., "Differences in Difficulty Among the CID W-22 Auditory Tests." Laryngoscope, 70:1560-5, 1960.

Elpern, B.S., "The Relative Stability of Half-List and Full-List Discrimination Tests." Laryngoscope, 71:30-6, 1961.

Fairbanks, G., Voice and Articulation Drillbook. New York: Harper Brothers, 1940.

Fastl, H., "Speech Intelligibility Tests with a Vocoder Based on the Hearing Sensation Sharpness." Acustica (Germany), 51(2):99-102, 1982.

Geffner, D., and Donovan, N., "Intelligibility Functions of Normal and Sensorineural Loss Subjects on the W-22 Lists." JAR, 14:82-7, 1974.

Ginzel, A., Pedersen, C.B., Spliid, P.E., and Andersen, E., "The Effect of Age and Hearing Loss on the Identification of Synthetic /b,d,g/-Stimuli." Scand. Aud., 11(2):103-12, 1982a.

Ginzel, A., Pedersen, C.B., Spliid, P.E., and Andersen, E., "The Role of Temporal Factors in Auditory Perception of Consonants and Vowels." Scand. Aud., 11(2):93-100, 1982b.

Godfrey, J.J., and Millay, K.K., "Perception of Synthetic Speech Sounds by Hearing-Impaired Listeners." JAR, 20(3):187-203, 1980.

Goetzinger, C.P., "Word Discrimination Testing." IN: Handbook of Clinical Audiology, Katz, J., (Ed.), Baltimore: The Williams and Wilkins Company, 1978.

Gold B., and Tierney, J., "LPC Synthesis with Simplified Spectrum Flattening of the Excitation Function." JASA, 72(4):1306-9, 1982.

Hirsh, I.J., Davis, H., Silverman, S.R., Reynolds, E.G., Eldert, E., and Benson, R.W., "Development of Materials for Speech Audiometry." JSHD, 17:321-37, 1952.

Hirsh, I.J., Reynolds, E.G., and Joseph, M., "Intelligibility of Different Speech Materials." JASA, 26(4):530-8, 1954.

Hodgson, W.R., Basic Audiologic Evaluation. Baltimore: The Williams and Wilkins Company, 1980.

Hood, J.D., and Poole, J.P., "Influence of the Speaker and Other Factors Affecting Speech Intelligibility." Audiology, 19:434-55, 1980.

Ingemann, F., "Speech Synthesis by Rule Using the FOVE Program." Haskins Laboratories Status Report on Speech Research, SR-54:165-73, 1978.

Ingemann, F., "The Contribution of Natural Durations to Speech Synthesized by FOVE Rules." Haskins Laboratories Status Report on Speech Research, SR-58:177-84, 1979.

Jenkins, J.J., and Franklin, L.D., "Recall of Passages of Synthetic Speech." Bulletin of the Psychonomic Society, 20(4):203-6, 1982.

Jerger, J., Speaks, C., and Trammell, J.L., "A New Approach to Speech Audiometry." JSHD, 33:318-29, 1968.

Kamm, C., Carterette, E.C., Morgan, D.E., and Dirks, D.D., "Use of Digitized Speech Materials in Audiological Research." JSHR, 23(4):709-21, 1980.

Kitawaki, N., Itoh, K., and Kakehi, K., "Speech Quality Measurement Methods for Synthesized Speech." Rev. Electr. Commun. Lab. (Japan), 29(9-10):895-906, 1981.

Kiukaanniemi, H.J., and Mattila, P., "Long-Term Speech Spectra." Scand. Aud., 9:67-72, 1980.

Klatt, D.H., "Word Verification in a Speech Understanding System." IN: Speech Recognition: Invited Papers Presented at the 1974 IEEE Symposium, Reddy, D.R., (Ed.), New York: Academic Press, 1975.

Klatt, D.H., "Structure of a Phonological Rule Component for a Synthesis-by-Rule Program." IEEE Trans. Acoust., Speech and Signal Process., ASSP-24:391-8, 1976.

Klatt, D.H., "Software for a Cascade/Parallel Formant Synthesizer." JASA, 67(3):971-95, 1980.

Kreul, E., Bell, D., and Nixon, J., "Factors Affecting Speech Discrimination Test Difficulty." JSHR, 12:281-7, 1969.

Lehiste, I., and Pederson, G.E., "Linguistic Considerations in the Study of Speech Intelligibility." JASA, 31(3):280-6, 1959.

Leng, G.W., Seng, L.W., and Kee, L.W., "A Study in Robotics-Teaching a Robot to Speak." IES J. (Singapore), 21(2):61-88, 1981.

Lerner, E.J., "Products that Talk." IEEE Spectrum, 19(7):32-7, 1982.

Levin, R.L., "The Intelligibility of Different Kinds of Test Materials Used in Speech Audiometry." M.A. Thesis, Washington University, 1952.

Levinson, E., and Kraat, A., "Intelligibility in Two Synthetic Speech Systems: A Pilot Study." Working Papers in Speech-Language Pathology and Audiology, Twelfth Volume of the reports of studies for the Speech and Hearing Center, Queens College of the City University of New York, 1984.

Levitt, H., "Computer Applications in Audiology and Rehabilitation of the Hearing Impaired." J. Commun. Disord.,13(6):471-81, 1980.

Liberman, A.M., Ingemann, F., Lisker, L., Delattre, P., and Cooper, F., "Minimal Rules for Synthesizing Speech." JASA, 31(11):1490-9, 1959.

Lindblom, B., "Spectrographic Study of Vowel Reduction." JASA, 35(11):1773-81, 1963.

Lowe, S.S., Cullen, J.K., Berlin, C.I., Thompson, C.L., and Willett, M.E., "Perception of Simultaneous Dichotic and Monotic Monosyllables." JSHR, 13(4):812-22, 1970.

Lundborg, T., Rosenhamer, H., Murray, T., and Zwetnow, N., "Information Abundance of Speech and Distorted Speech Testing in Topical Diagnosis Within the C.N.S." Scand. Aud., 4:9-16, 1975.

Lynn, J.M., and Brotman, S.R., "Perceptual Significance of the CID W-22 Carrier Phrase." Ear and Hearing, 2(3):95-9, 1981.

Martin, F.N., and Forbis, N.K., "The Present Status of Audiometric Practice: A Follow-Up Study." ASHA, 20:531-41, 1978.

Mattingly, I.G., "Experimental Methods for Speech Synthesis by Rule." IEEE Trans. Audio. Electroacoust., 16:198-202, 1968.

Mattingly, I.G., "Phonetic Representation and Speech Synthesis by Rule." Haskins Laboratories Status Report on Speech Research, SR-61:15-21, 1980.

Merzenich, M.M., Byers, C.L., White, M., and Vivion, M.C., "Cochlear Implant Prostheses: Strategies and Progress." Ann. Biomed. Eng., 8(4-6):361-8, 1980.

Miller, G.A., Heise, G.A., and Lichten, W., "The Intelligibility of Speech as a Function of the Context of the Test Materials." J. Exp. Psychol., 41:329-35, 1951.

Miller, G.A., and Nicely, P.E., "An Analysis of Perceptual Confusions Among Some English Consonants." JASA, 27(2):338-52, 1955.

Mitchell, P.D., "Test of Differentiation of Phonemic Feature Contrasts." JASA, Suppl. 55:S55, 1974.

Morgan, N., Talking Chips. New York: McGraw-Hill, Inc., 1984.

Nakatsui, M., and Mermelstein, P., "Subjective Speech-to-Noise Ratio as a Measure of Speech Quality for Digital Waveform Coders." JASA, 72(4):1136-44, 1982.

Nelson, D.A., and Chaiklin, J.B., "Writedown Versus Talkback Scoring and Scoring Bias in Speech Discrimination Testing." JSHR, 13(3):645-54, 1970.

Nye, P.W., and Gaitenby, J.H., "Consonant Intelligibility in Synthetic Speech and in a Natural Speech Control (Modified Rhyme Test Results)." Haskins Laboratories Status Report on Speech Research, SR-33:77-91, 1973.

Nye, P.W., and Gaitenby, J.H., "The Intelligibility of Synthetic Monosyllable Words in Short, Syntactically Normal Sentences." Haskins Laboratories Status Report on Speech Research, SR-37/38:169-90, 1974.

Nye, P.W., Hankins, J.D., Rand, T., Mattingly, I.G., and Cooper, F.S., "A Plan for the Field Evaluation of an Automated Reading System for the Blind." IEEE Trans. Audio. Electroacoust., AU-21(3):265-8, 1973.

Nye, P.W., Ingemann, F., and Donald, L., "Synthetic Speech Comprehension: A Comparison of Listener Performances with and Preferences Among Different Speech Forms." Haskins Laboratories Status Report on Speech Research, SR-41:117-25, 1975.

Oggerino, J.J., "An Evaluation of a Talking Machine: The HC 120 Phonic Mirror Handivoice." M.S. Thesis, Portland State University, 1980.

Paliwal, K.K., and Rao, P.V.S., "Synthesis-Based Recognition of Continuous Speech." JASA, 71(4):1016-24, 1982.

Palmer, J.M., "The Effect of Speaker Differences on the Intelligibility of Phonetically Balanced Word Lists." JSHD, 20(2):192-5, 1955.

Peterson, G.E., and Lehiste, I., "Revised CNC Lists for Auditory Tests." JSHD, 27:62-70, 1962.

Pisoni, D.B., "Variability of Vowel Formant Frequencies and the Quantal Theory of Speech: A First Report." Phonetica, 37(5-6):285-305, 1980.

Pisoni, D.B., "Perception of Speech: The Human Listener as a Cognitive Interface." Speech Technol., 1(2):10-24, 1982.

Pisoni, D.B., and Hunnicutt, S., "Perceptual Evaluation of MITALK: The MIT Unrestricted Text-to-Speech System." Paper presented at the Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing. Denver, Co., pp. 572-5, 1980.

Prado, E., "Voice Input for CAD/CAM." Computer Graphics World, 111-3, June, 1983.

Rahko, T., Karjalainen, M.A., Laine, U.K., and Lavonen, S., "Speech Audiometry by a Speech Synthesizer." Arch Otorhinolaryngol., 222:85-9, 1979.

Rahko, T., Karjalainen, M., Laine, U., and Lavonen, A., "Effect of Word Speed, Number of Words, and Habituation on the Discrimination Score of Synthesized Speech." Ann. Otol. Rhinol. Laryngol., 89:69-71, 1980.

Rao, P.V.S., and Thosar, R.B., "A Programming System for Studies in Speech Synthesis." IEEE Trans. Acoust., Speech and Signal Process., ASSP-22(3):217-25, 1974.

Rintelmann, W.F., Schumaier, D.R., Jetty, A.J., Burchfield, S.A., Beasley, D.S., Mosher, N.A., Mosher, R.A., and Penley, E.D., "Six Experiments on Speech Discrimination Utilizing CNC Monosyllables." JAR, Suppl. 2:1-30, 1974.

Ross, M., and Huntington, D.A., "Concerning the Reliability and Equivalency of the CID W-22 Auditory Tests." JAR, 2:220-8, 1962.

Sanders, D.A., Aural Rehabilitation. Englewood Cliffs, New Jersey: Prentice-Hall, Inc., 1971.

Sargent, D.C., "Rhythmic Cues Aid Lip Readers." IEEE Spectrum, 19(4):46-9, 1982.

Schultz, M.C., "Suggested Improvements in Speech Discrimination Testing." JAR, 4(1):1-14, 1964.

Schwartz, A.H., and Goldman, R., "Variables Influencing Performance on Speech-Sound Discrimination Tests." JSHR, 17:25-32, 1974.

Silverman, S.R., and Hirsh, I.J., "Problems Related to the Use of Speech in Clinical Audiometry." Ann. Otol. Rhinol. Laryngol., 64:1234-44, 1955.

Snow, J.B., Rintelmann, W.F., Miller, J.M., and Konkle, D.F., "Central Auditory Imperception." Laryngoscope, 87:1450-71, 1977.

Stevens, J.H., "Monosyllabic Speech Tests." IN: Handbook of Clinical Audiology, Katz, J., (Ed.), Baltimore: The Williams and Wilkins Company, 1978.

Street Electronics Corporation, Echo II Speech Synthesizer. California: Street Electronics Corporation, 1982.

Teja, E.R., "Voice-Output Units Show Improvement in Speech Quality, Vocabulary, Price." EDN, 27(11):71-80, 1982.

Tillman, T.W., and Jerger, J.F., "Some Factors Affecting the Spondee Threshold in Normal-Hearing Subjects." JSHR, 2(2):141-6, 1959.

Tillman, T.W., and Olsen, W.O., "Speech Audiometry." IN: Modern Developments in Audiology, Jerger, J., (Ed.), New York: Academic Press, 1973.

Ventry, I.M., Chaiklin, J.B., and Dixon, R.F., (Ed.), Hearing Measurement: A Book of Readings. New York: Meredith Corporation, 1971.

Voiers, W.D., "Evaluating Processed Speech Using the Diagnostic Rhyme Test." Speech Technol., 2(4):30-9, 1983.

Williams, D.H., Simpson, C.A., and Nordinger, C., "Comparative Intelligibility of VPSP Text-to-Phoneme Speech and Handivoice Pre-Stored and Phoneme Speech." PLRA Research Report, 80-2:1-6, 1980.

Williams, J.M., "Speech Output Helps Disabled People Lead Productive Lives." Speech Technol., 1(3):69-72, 1982.

Winer, B.J., Statistical Principles in Experimental Design. New York: McGraw-Hill, Inc., 1962.

Wise, C.M., Applied Phonetics. Englewood Cliffs, New Jersey: Prentice-Hall, Inc., 1957.

Wong, D.Y., and Markel, J.D., "An Intelligibility Evaluation of Several Linear Prediction Vocoder Modifications." IEEE Trans. Acoust., Speech and Signal Process., ASSP-26(5):424-35, 1978.

Yorkston, K., and Beukelman, D., Assessment of Intelligibility of Dysarthric Speech. Seattle: C.C. Publications, Inc., 1981.

Zahorian, S.A., and Rothenberg, M., "Principal-Components Analysis for Low-Redundancy Encoding of Speech Spectra." JASA, 69(3):832-45, 1981.

APPENDIX A

SPEAKEASY COMMANDS

All phoneme words must be preceded by a CTRL-V.
PITCH and RATE commands are the same as in TEXTALKER.

SPEAKEASY PHONEME CODES

VOWELS

cat	a	A
lot	o	;
caught	ô	*
let	e	E
see	ē	&
hid	i	I
book	oo	Q
but	u	U
due	o͞o	:
about	de	'

DIPHTHONGS

cake	ā	@
tie	ī	!
toe	ō	O
pound	ou	#
toil	ôi	?
you	ū	%

"R" COLORED VOWELS

car	är	;R
chair	âr, er	@R
her	ür, ĕr	'R
hear	ēr	&R
fire	īr	!R
for	ôr	OR
tour	oor	QR
hour	our	#R

VOICED CONSONANTS

let	l	L
many	m	M
no	n	N
sing	ŋ	/
red	r	R
this	*th*	(
very	v	V
wet	w	W
yes	y	Y
zero	z	Z
azure	zh	X

STOP CONSONANTS

bat	b	B
dog	d	D
get	g	G
kick	k	K
pet	p	P
tie	t	T
check	ch	C
job	j	J

UNVOICED FRICATIVES

fit	f	F
hat	h	H
see	s	S
she	sh	$
think	th	)

INFLECTION

STRESSED 3 **NORMAL 2** **REDUCED 1** **SCHWA 0**

PITCH

RISING > **FLAT -** **FALLING <** **PAUSE ,**

Street Electronics Corporation
1140 Mark Avenue
Carpinteria, California 93103
(805) 684-4593

APPENDIX B

ENCODED FORMS AND IPA SYMBOLS FOR NU-6 TEST WORDS SCRAMBLING A, USING THE TEXTALKER AND SPEAKEASY PROGRAMS

		WORD	TEXTALKER	SPEAKEASY	IPA
*	1	PICK	7PI2K	7PI2K	pɪk
	2	ROOM	7R:2M	7R:2>M	rum
	3	NICE	7N!2S	7N'3!2S	naɪs
*	4	SAID	7SE3D	7SE3D	sɛd
	5	FAIL	7F@3L	7FU17@3L	fel
	6	SOUTH	7S#2)	7S#3)	saʊθ
	7	WHITE	7W!2T	7HW!2T	(h)wait
	8	KEEP	7K&3P	9K&39P'Ø	kip
	9	DEAD	7DE3D	7DE3D'Ø	dɛd
*	10	LOAF	7LO2F	7LO2F	lof
	11	DAB	7DA2B	7DA2B'Ø	dæb
	12	NUMB	7NU2MB	7NNU2MMB	nʌm
	13	JUICE	7J:2!2S	7J:2S	dʒus
	14	CHIEF	7C&2F	7C&3F	tʃif
	15	MERGE	7M'R2J	7M'R2J'Ø	mɜd
	16	WAG	7WA2G	7WA2G'Ø	wæg
	17	RAIN	7R@3N	7RR@3NN	ren
	18	WITCH	7WI2C	7HWI2C	wɪtʃ
	19	SOAP	7SO2P	7SO2P'Ø	sop
	20	YOUNG	7YU3/	7YU39/	jʌŋ
	21	TON	7TU2N	7TU33N	tʌn
	22	KEG	7KE2G	7KE3G	kɛg
	23	CALM	7K*2LM	7K*2'Ø7LM	kɔ(l)m/kɑ(l)m
	24	TOOL	7T:2L	7T9<:3L	tul
	25	PIKE	7P!2K	7P<!2K	paɪk
	26	MILL	7MI2L	7MMI2LL	mɪl
	27	HUSH	7HU2$	7H'Ø9U2$	hʌʃ
	28	SHACK	7$A2K	7>$A2>K	ʃæk
	29	READ	7RE3D	7R&39D'Ø	rid (rɛd-TT)
	30	ROT	7R;2T	7<R;25T	rɑt
	31	HATE	7H@3T	7H'Ø@3T	het
	32	LIVE	7L!2V	7LI2VV	lɪv (laɪv-TT)
	33	BOOK	7BQ3K	7B'2Q3K	bʊk
	34	VOICE	7V?2S	7V?2SS	vɔɪs
	35	GAZE	7G@3ZS	7G'Ø7@2ZS	gez
*	36	PAD	7PA2D	7PA2D	pæd
	37	THOUGHT	7)*3T	7))*3T	θɔt
*	38	BOUGHT	7B*3T	7B*3T	bɔt
	39	TURN	7T'R2N	7T7'R2NN	tɜn
	40	CHAIR	7C@R3	7C3@R3	tʃɛr
	41	LORE	7LOR2	9LLOR	lor
	42	BITE	7B!2T	7B!3T'Ø	bait

*	43	HAZE	7H@3ZS	7H@3ZS	hez
*	44	MATCH	7MA2C	7MA2C	mætʃ
*	45	LEARN	7L'R3N	7L'R3N	lɜn
	46	SHAWL	7$*3L	7$$7*3LL	ʃɔl
	47	DEEP	7D&3P	7D<&3P	dip
	48	GIN	7JI2N	7JI2NN	dʒɪn
	49	GOAL	7GO2L	7GO17L	gol
	50	FAR	7F;R3	7F;17;R3	far

* = Words encoded the same for both TEXTALKER and SPEAKEASY.